Scholarly Article

Journal of Service Research
1-27

DOI: 10.1177/10946705209751 43
journals.sagepub.com/home/jsr


Harvesting Online Reviews to Identify the Competitor Set in a Service Business: Evidence From the Hotel Industry

Fei Ye[1], Qian Xia[1,2], Minhao Zhang[3], Yuanzhu Zhan[4], and Yina Li[1]

Abstract

In today's global service industry, online reviews posted by consumers offer critical information that influences subsequent consumers' purchasing decisions and firms' operation strategies. However, little research has been done on how the same information can be used to identify key competitors and improve services to increase competitiveness. In this article, we propose an analytical framework based on an improved *k*-nearest neighbor model and a latent Dirichlet allocation model for service managers to harvest online reviews to identify their key competitors and to evaluate the strengths and weaknesses of their businesses. With a sample comprising over 8 million customer reviews of 6,409 hotels in 50 Chinese cities from Ctrip.com, we validate the effectiveness of the proposed approach in the analysis of a hotel's service competitiveness and its key competitors. The findings indicate that the importance of particular attributes of a hotel varies in different segments according to hotel star ratings. This study extends the literature by bridging online reviews and competitor identification for service industries. It also contributes to practice by offering a systematic and effective way for managers to identify their key competitors, monitor market preferences, ensure service quality, and formulate effective marketing strategies.

Keywords

online reviews, competitor identification, *k*-nearest neighbor, latent Dirichlet allocation, hotel attributes

Online reviews generated by consumers are becoming increasingly influential in today's rapidly changing service business and particularly in the hotel industry (Mathwick and Mosteller 2017; Y. Wang et al. 2020; L. Wu et al. 2016). This is driven by the trends of globalization, aging populations, reduced travel costs, and increased leisure time—the service-intensive hotel industry has witnessed corresponding rising demand (Mohammed, Guillet, and Law 2014). Moreover, due to the evolution of Web 2.0, the number of hotel reviews posted on the websites of online travel agents (OTAs) such as Booking.com and TripAdvisor.com has grown enormously (K. Lu and Elwalda 2016). Recent market research has shown that over 49% of travelers will not choose a hotel without reviewing online comments (World Travel Market 2014), and approximately 35% of consumers modify their schedules after checking posts on OTAs (L. Wu et al. 2016). In addition, online reviews can have an important influence on service organizations' bottom line. For instance, Mathwick and Mosteller (2017) reported that a 1% improvement in online reputation could result in a 1.4% growth in revenue per hotel room.

Today, online reviews enable consumers to share their experiences and opinions at an unprecedented scale and speed. Such reviews present a substantial amount of rich information on competitors, particularly in the form of service comparisons (W. Wang, Yi, and Dai 2018). Although online reviews have been adopted throughout all areas of both service and manufacturing industries, the information included tends to be incredibly valuable for service industries (Mathwick and Mosteller 2017; L. Wu et al. 2016). Compared to physical products that typically have multiple features that can be easily classified and evaluated, the measurements that constitute "excellent" or "terrible" services tend to be complicated to objectively identify and define (Mankad et al. 2016). As a result, the subjective customer opinions that are embedded in online reviews become much more informative by comparison. Notably, the use of online reviews works for all service sectors, and managers today need to monitor and analyze both negative and positive online reviews in order to track the products,

[1] School of Business Administration, South China University of Technology, Guangzhou, China
[2] College of Business, Guizhou Minzu University, Guiyang, China
[3] Department of Management, University of Bristol, United Kingdom
[4] Management School, University of Liverpool, United Kingdom

Corresponding Author:
Yuanzhu Zhan, Management School, University of Liverpool, Chatham Street, Liverpool L69 7ZH, United Kingdom.
Email: yuanzhu.zhan@liverpool.ac.uk

services, promotions, and sales offered by their competitors (Jin, Ji, and Gu 2016). Pelsmacker, Tilburg, and Holthof (2018) highlight that the volume and valence of online reviews reflect the competitive marketing strategies of service providers and can have an effect on their market performance. Therefore, it is of great importance to develop an approach to support the analysis of the competitiveness of a service provider and the identification of its key competitors by using online reviews.

A comprehensive literature review shows that research bridging online reviews and competitor identification in service research is in its infancy, as there is a lack of operational approaches to extend the scope of either area. On the one hand, prior studies have revealed the use and effects of online reviews in various fields, such as marketing (K. Lu and Elwalda 2016; Pelsmacker, Tilburg, and Holthof 2018; Ye, Law, and Gu 2009), information systems research (Chen and Yao 2016; Filieri et al. 2018; Mariani, Borghi, and Gretzel 2019), and innovation management (Algesheimer et al. 2011; K. Lu and Elwalda 2016; Moe and Trusov 2011; Zhan et al. 2020). However, none of these studies takes the perspective of service providers, and so they do not advance current discourse in relation to identifying key competitors and improving services. Moreover, most of these studies consider only a limited amount of information from online reviews (e.g., they might extract a single summary opinion from a review), which cannot provide managers with an integrated and comprehensive set of competitors. On the other hand, the service literature has established that consumer evaluations of a service are greatly affected by interactions among consumers, operational approaches, information systems, staff, and companies (Brown and Dev 2000). These factors have been studied in relation to service quality, service representatives, and service blueprinting (Holloway and Beatty 2003; Rapp et al. 2015; Tsai and Lu 2006), suggesting that services involve companies and customers in co-creation (Holloway and Beatty 2003; Kumar et al. 2010). In spite of their theoretical and practical implications, these factors have largely been overlooked in the operationalization of models that can identify competitor sets and harvest the value of online customer reviews (Antons and Breidbach 2018). Notably, an analytical framework is required that can integrate relevant attributes of a service and help companies analyze online customer reviews and identify their key competitors. Accordingly, research has increasingly suggested that new approaches, such as data analytics and machine-learning methods, are needed to improve service systems (Gur and Greckhamer 2019; Jin, Ji, and Gu 2016). This study argues that insights into the competitor set are more likely to be captured in rich online reviews than through company-based questionnaires. Therefore, the lack of an analytical framework centering on the identification of the competitor set is a critical oversight.

The main objective of this study is to develop an analytical approach to help managers harvest information from online reviews that will allow them to identify their competitor set. The study setting is the hotel industry, but the approach could be used by service companies in general. It is based on the integration of an improved *k*-nearest neighbor (*k*NN) model and a latent Dirichlet Allocation (LDA) model. Competitors are identified from online customer reviews, combined with hotel description data and online search ranking data. Although the online review data are highly important in determining a hotel's strategic plan, surprisingly few studies in the field of service research have utilized online customer reviews to identify competitors and, in practice, managers do not have systematic guidance on how best to process the vast amount of data present in online reviews (Antons and Breidbach 2018; Rapp et al. 2015). In light of this, the study proposes an integrated analytical approach that draws from a variety of disciplines (e.g., statistics, machine learning, and computer science) to conduct an in-depth analysis of online reviews to determine the importance of hotel attributes in different market segments (according to hotel star ratings).

This research makes three key contributions to the literature and practice. First, the service attributes identified from consumers' online reviews can support hotel managers in evaluating their perceived quality of services and their competitive environment. Importantly, those attributes depend on the market segment served by a particular hotel. Second, as online reviews normally include information on competitors, we propose a more effective analytical framework, based on a set of machine-learning techniques, for service managers to determine their key competitors and to identify their own company's weaknesses and strengths. This will, in turn, allow them to develop appropriate marketing strategies and make appropriate service improvements. Third, the proposed framework offers the opportunity for real-time analysis of the competitor set by applying analytical techniques. That is, it enables managers to conduct dynamic analysis to monitor their key competitors and changes to the market environment by applying up-to-date information from consumers' online reviews.

Literature Review

Two fields of the literature relate to the present study: the use of Online Reviews for Value Co-Creation and Service Improvement subsection and Competitor Identification in the Service Domain subsection. Also, the existing methods and approaches for competitor identification are compared in subsection Methods and Approaches for Competitor Identification, and the settings for research regarding customers' hotel selection via OTAs are presented in subsection Research Settings: Customer Hotel Selection via OTAs.

Online Reviews for Value Co-Creation and Service Improvement

The impact of consumer-company interactions on consumer evaluation of a service has long been seen as the process nature of services in the literature (Antons and Breidbach 2018; Brown and Dev 2000; Parasuraman, Berry, and Zeithaml 1993). Identifying these interactions can help companies to enhance their understanding of the "customer encounter,"

which is defined as a customer's direct interactions with the service during a specific period (Ordenes et al. 2014). Studies show that the encounters are important for consumers' evaluation of service quality (Parasuraman, Berry, and Zeithaml 1993), customer loyalty (Brodie et al. 2011; Kumar et al. 2010), and customer satisfaction (Algesheimer et al. 2011; Nasution and Mavondo 2008). According to L. Wu et al. (2016), key encounters between companies and consumers can happen in different ways, such as face-to-face interactions, telephone, email, and the internet. To increase the quality of these encounters, the literature has studied the co-creative nature of services and consumers' evaluations are treated as the outcomes of the multiple activities provided and resources applied during the service (Kumar and Pansari 2016; Mathwick and Mosteller 2017).

Moreover, the service literature summarizes critical factors in the service process that enhance consumers' realization of value (Filieri et al. 2018; Nasution and Mavondo 2008). When receiving services, consumers combine activities offered by the company with external resources and use various approaches to generate value for themselves (A. C. C. Lu, Gursoy, and Lu 2016; Ordenes et al. 2014). During the consumer-company interactions, consumers' value creation can be affected by the information platform (e.g., online forums and communities) provided by the companies (Antons and Breidbach 2018; Thakur 2018). These value co-creation platforms offer both the company and the consumer access to information that enables various activities, and different results are possible based on how the interaction proceeds. Companies work as value facilitators who support consumers in their value creation by offering them the necessary information and resources (Ordenes et al. 2014).

To facilitate the value co-creation and offer the right services to consumers (Kumar et al. 2010), it is important for companies to gain insights into consumers' evaluations of their experiences and their perception of the value of the company's services in a context defined by the consumers (Gao et al. 2018; Parasuraman, Berry, and Zeithaml 1993). This can be achieved by companies via harvesting online customer reviews through information platforms during or following interactions (Gur and Greckhamer 2019; Jin, Ji, and Gu 2016). According to Tan et al. (2018), although online reviews cannot directly lead to value generation for companies, they can result in internal process development and actionable information for decision making if proper analytical approaches are in place. For example, an analytical approach can be developed to enable managers to collect all their online reviews and other sources of information on their interactions with consumers, so that they can evaluate the competitive environment effectively and respond to consumers' feedback in a timely manner (W. Wang, Yi, and Dai 2018). Also, the company's weaknesses and strengths can be evaluated, which in turn will allow managers to develop appropriate marketing strategies and service improvements. Moreover, the analysis of online information can be done much more quickly and with information that is much more up to date than could be done using traditional means (Antons and Breidbach 2018). Nonetheless, studies have been rare in the service literature that systematically investigate the value co-creation process by harvesting the value of online reviews.

Competitor Identification in the Service Domain

Competition in the service industries is widely regarded as complex and dynamic (Du, Hu, and Damangir 2015; Nam, Joshi, and Kannan 2017). To identify competitors, managers normally focus on a small group of companies because of their bounded rationality and limited managerial resources (Peteraf and Bergen 2003). This approach is in line with the cognitive categorization view, which explains why companies identify only simple competitor sets and pay attention to just a few categories of business rivals (Baum and Lant 2003; Hatzijordanou, Bohn, and Terzidis 2019).

Service competitors can be defined in different ways. According to Ng, Westgren, and Sonka (2009), competition can be interpreted differently by stakeholders within a value chain, as they may have different perceptions of rivals. Most of the literature uses the concept of service substitution to define competitors, whereby service attributes (e.g., pricing and service cycle time) are compared to identify which other service providers are most similar (e.g., Clark and Montgomery 1999). The important service attributes classified by the early research build a strong foundation of understanding the dimensions on which competitors are best defined. For example, by identifying a company's service shortfalls and strengths, SERVQUAL can help to define the competitor set and to determine which competitors share particular disadvantages and advantages attributes (Brown and Dev 2000; Parasuraman, Berry, and Zeithaml 1993). In this way, managers are recommended to evaluate the strengths and weaknesses of the company as well as those of its competitors through predefined scales and measurements (Clark and Montgomery 1999). However, this approach to competitor identification has been criticized for its subjective bias. Ng, Westgren, and Sonka (2009) suggest that when interpreting competition, managers may have different "blind spots" due to their personal characteristics and experiences.

The literature suggests that service improvement begins by comparing what consumers believe a firm ought to provide with what they perceive the firm's actual service to be (Antons et al. 2018; Brown and Dev 2000; Gao et al. 2018). Accordingly, the competitors of service providers can be defined from the customer perspective by benchmarking the market preference for particular service attributes (Baum and Lant 2003; Sidhu, Nijssen, and Commandeur 2000). In other words, it aims to contribute to the knowledge regarding how customers define the competitor set for a focal firm. This approach is consistent with the view of service demand, which defines competitors as all the companies that aim to meet a similar set of customer demands. According to Sidhu, Nijssen, and Commandeur (2000), in comparison with other perspectives, the customer perspective normally identifies a wider and larger competitor

set (i.e., direct and immediate competitors, as well as potential competitors), which may even span different industries, and so competitive boundaries are blurry. Identifying competitors from the customer perspective can, nevertheless, reduce the adverse effects of short-sightedness, competition asymmetry, and the "competitive blind spots" of managers (J. B. Kim, Albuquerque, and Bronnenberg 2011; Ng, Westgren, and Sonka 2009).

In addition, how to evaluate competitors is another important topic within the growing body of service literature. The superiority of relative service metrics over absolute measures of satisfaction is emphasized (e.g., Keiningham, Buoye, and Ball 2015). Recent studies show that instead of using a numerical value to measure customer satisfaction (i.e., the absolute measures), a ranking of competitors according to customer satisfaction is found to be more strongly associated with their share of wallet (Keiningham et al. 2015). According to Keiningham et al. (2014), although the use of relative metrics is robust in measuring service success, the question of "relative to whom" might be challenging. Therefore, developing an effective approach to competitor identification is a critical preliminary step in understanding customers' perceptions of and attitudes to service providers.

Methods and Approaches for Competitor Identification

To identify competitors in the service domain, researchers have applied different methods and approaches in terms of the nature of the data resources and expertise required. Two commonly used methods have been surveys and archival studies.

According to Gur and Greckhamer (2019), quantitative empirical studies using cross-sectional survey data are the most common method for identifying competitors. Such research has generally taken a company perspective or a customer perspective (Gur and Greckhamer 2019). On the one hand, national or international statistical data have been widely utilized in empirical studies of competitor identification (Cooper and Inoue 1996; J. Wu and Olk 2014). On the other hand, researchers have explored how companies satisfy customer needs and analyzed customers' brand-switching behaviors (DeSarbo and Grewal 2007; Wieringa and Verhoef 2007). The study of brand switching (e.g., Roos, Edvardsson, and Gustafsson 2004; Wieringa and Verhoef 2007) normally employs behavioral panel data in which one observes brand switching, for example, in examining customer perceptions of brands, and may apply a log-linear modeling framework to investigate which brands have similar image profiles to identify whether they form sharing or switching partitions.

The second method is the archival study. The early research analyzing firms' archival data to identify their competitors focused on defining strategic groups of firms that share certain characteristics such as strategies, resources, and environment (Peteraf and Bergen 2003). In the domain of service research, the banking sector was one of the earliest "laboratories" for researchers using archival data to identify competitors through the analysis of strategic groups (Amel and Rhoades 1988). Various approaches have been applied to identify the strategic groups so as to find closely competing companies within the same industry. However, the main issue with this method is that the structure and boundaries of these strategic groups are usually ambiguous (Baum and Lant 2003). In addition, the method has typically employed a two-step approach. The researchers first apply factor analysis to identify the underlying dimensions and then they identify the strategic group using cluster analysis. As a result, multidimensional scaling (MDS) has been widely adopted to overcome the weaknesses of the cluster analysis such as the inconsistency of the factors that emerge and the overlooking of the time factor (DeSarbo and Grewal 2007).

To benchmark our proposed method against relevant studies, we summarize previous approaches for mapping and analyzing competitive market structures in Table 1. Previous studies have depended on data captured from questionnaires and surveys. For example, Cooper and Inoue (1996) apply archival data (questionnaires collected by Rogers National Research) to determine the preferences of different customer segments while DeSarbo and Grewal (2007) use survey data to investigate purchase intentions for vehicles through an asymmetric MDS approach.

With the development of information technologies, service companies today are paying more attention to understanding their competition from a customer's perspective given the large volume, variety, and veracity of user-generated content. J. B. Kim, Albuquerque, and Bronnenberg (2011) extracted data from Amazon.com on customer search patterns. Netzer et al. (2012) used user-generated textual data from an online automobile forum to identify competitive market structures. Du, Hu, and Damangir (2015) combined sales data from Automotive News together with search trends from Google Trends to illustrate evolving customer preferences. Nam, Joshi, and Kannan (2017) aggregated textual data from a social tagging platform to identify user-generated social tags. Additionally, the development of competitor sets is related to a subset of product attributes. Studies show that product attributes beyond marketers' control can changes customers' buying decisions (Baum and Lant 2003; Du, Hu, and Damangir 2015; J. B. Kim, Albuquerque, and Bronnenberg 2011).

Although previous approaches for identifying competitors have their merits, there are some challenges related to basic assumptions, data availability, and the visualization of large product categories. First of all, previous studies such as Cooper and Inoue (1996) and DeSarbo and Grewal (2007) collected their data through questionnaires and surveys. However, these data limit the potential to study consumer durables in large markets involving thousands of products. For example, customers are not likely to buy durable goods (e.g., vehicles or household appliances) very often. Therefore, these approaches are bounded by the cognitive capacity of customers. According to Ringel and Skiera (2016), even when studying just a handful of alternative products that customers tend to consider at the same time, it is questionable whether respondents can appropriately recall previous buying decisions or predict future purchase intent. Also, questionnaires and surveys tend to be

Table 1. Comparison of Studies Analyzing and Mapping Competitive Market Structure.

	Cooper and Inoue (1996)	DeSarbo and Grewal (2007)	J. B. Kim, Albuquerque, and Bronnenberg (2011)	Netzer et al. (2012)	Du, Hu, and Damangir (2015)	Nam, Joshi, and Kannan (2017)	This Study
Objective	Analyze market structures by determining the preferences of different customer segments	Identify and represent asymmetric competitive market structure	Propose an approach to visualize rich consumer search patterns	Convert the user-generated content to market structures and competitive insights	Propose a market response model to leverage trends in online searches in evolving customer preferences	Analyze the set of brand associations obtained from user-generated social tags	An approach to identify competitors and visualize the competitive landscape
Source	Rogers National Research and Consumer Reports	Survey	Amazon.com	Online forum (Edmunds.com)	Automotive News and Google Trends	A social tagging platform	Online travel agent platform (Ctrip.com)
Type of data	Questionnaire about 106 different car models	Two different industries and 10 products	Search data on four companies and 62 products	Textual data on 30 car brands and 169 products	Sales data and search trends for 80 nonluxury vehicles	Textual data on seven brands	Numerical data and textual data on 6,409 hotels
Approach	A competitive market-structure model	Asymmetric multidimensional scaling (MDS)	Hierarchical MDS	Text mining and net analysis approaches (classical MDS)	A log-log sales response model	Text mining and data reduction techniques	An improved *k*-nearest neighbor model and a latent Dirichlet allocation model
Nondurables	—	—	—	—	—	Yes	Yes
Low cost	—	—	Yes	Yes	Yes	Yes	Yes
Real time	—	—	Yes	Yes	Yes	Yes	Yes
Exploring key factors	—	—	Yes	—	Yes	Yes	Yes
Large product categories	—	—	—	—	—	—	Yes

time-consuming to complete and costly to administer and cannot be used to indicate real-time customer behaviors (J. B. Kim, Albuquerque, and Bronnenberg 2011; Nam, Joshi, and Kannan 2017; Ringel and Skiera 2016). Besides, the models developed by Cooper and Inoue (1996) and Du, Hu, and Damangir (2015) were based on a number of mathematical assumptions and there were ambiguities regarding consumer search intentions. Thus, the methods and results may not be fully applicable to companies in practice.

For a market involving a small number of products, it is relatively simple to demonstrate the competitive market structure by presenting dots on an XY graph, where each dot indicates a different product. However, as the number of products increases, the graphical presentation rapidly becomes a dense clump of dots, making it difficult to interpret the results (J. B. Kim, Albuquerque, and Bronnenberg 2011; Ringel and Skiera 2016). Although an additional dimension can be applied to mitigate this effect (DeSarbo and Grewal 2007), this should be avoided wherever possible because it tends to be difficult to check and explain the results (Ringel and Skiera 2016). Meanwhile, the selection of similarity measures can be ambiguous but can play an important part in the analysis. Appendix E shows some examples of the MDS maps generated using different similarity measures (we further explain this in Model Evaluation section).

Furthermore, MDS techniques are especially sensitive to the size of the data set being analyzed (Moore and Holbrook 1982; Ringel and Skiera 2016). It is inherent to the technique that the accuracy of data positions deteriorates when the data set becomes large (Buja et al. 2008). Issues such as the circular bending effect are common in MDS analysis (Carroll and Arabie 1980). This can result in inaccurate identification of competitive structures and, in particular, competitive relationships can be shown to be tighter than they actually are (Diaconis,

Goel, and Holmes 2008; Moore and Holbrook 1982; Ringel and Skiera 2016). Nonetheless, our proposed method enables companies to identify their competitors effectively even for large product categories—that is, categories containing over 6,000 products. A comprehensive competitive map is created which enables companies to conduct real-time analysis of their competitor sets with consideration of specific strengths and weaknesses.

Research Settings: Customer Hotel Selection via OTAs

Like many other labor-intensive service industries, the hotel industry is under increasing pressure (e.g., to lower its costs and offer more high-quality services) and is highly concerned with competitor identification (Brown and Dev 2000; Kim and Canina 2011; Mohammed, Guillet, and Law 2014). Competitor identification is, therefore, a vital initial step in market evaluation, service improvement, and strategy development (J. Y. Kim and Canina 2011). As information search is often customers' initial step, at which companies can affect their decision making, it is important to understand how online customers select hotels, especially in the era of big data. According to A. C. C. Lu, Gursoy, and Lu (2016), customers today want to compare products on different attributes before making their decisions. While a tremendous amount of external information is available to customers, they tend to use a small number of hotel attributes in their prepurchase information search. Previous studies have found that the importance customers attach to particular types of information in their prepurchase search depends on, for example, situational factors (e.g., risk perceptions and previous experience), product characteristics (e.g., type of trip and destination type), decision complexity (e.g., number of alternatives), and consumer characteristics (e.g., educational level and culture; A. C. C. Lu, Gursoy, and Lu 2016; Tan et al. 2018).

The information search process is quite different through OTAs. According to the 2018 Chinese Travel Consumer Report,[1] over 77% of hotel bookings in China are made through OTA websites, and this figure increases to 81% for bookings made on a phone app. In the present study of hotel selection from the customers' perspective, data were obtained from Ctrip (www.Ctrip.com), a leading OTA that provides flight tickets, hotel reservations, and tourist resort products in China. There are two main reasons for using Ctrip.com. First, according to Shao and Kenney (2018), Ctrip has become one of the largest and fastest growing OTAs. It attracts over 135 million users in the Chinese market, and over the period 2017–2018 had a compound growth rate of 25%. Ctrip's 2018 annual report shows that its net income had reached US$4.5 billion for the full year of 2018, a 16% rise year on year. Second, unlike the data from other platforms, data generated from Ctrip.com can be considered "open" (Ctrip 2017), and researchers and organizations have used data from Ctrip.com to monitor and analyze challenging issues in diverse fields (Leung, Law, and Lee 2011; Shao and Kenney 2018; Ye, Law, and Gu 2009).

OTAs make the search process simple and effective. To access hotel information from OTAs, customers usually are required to enter some basic information such as their destinations and dates of check-in and check-out. To understand customers' requirements as well as avoid information overload, leading OTAs filter the hotels for customers based on certain predefined criteria, such as star rating, price range, and location. Star rating has been identified as the most important selection criterion for customers when they select a hotel via an OTA. On Ctrip.com, customers need to specify a hotel star rating (i.e., from two to five stars) before doing their initial search (as shown in Appendix D). Notably, the 2017 Ctrip Hotel White Paper shows that over 76% of the hotel searches on Ctrip.com were associated with a star rating (while 11.61% of the searches were price-related).[2] The report also identifies a rapidly increasing demand for highly rated hotels when customers were searching Ctrip's listed hotels by star rating. This, in turn, indicates that hotels listed on Ctrip.com are more likely to compete with each other within the same star rating.

Although we acknowledge that a variety of factors can affect potential customers' search process, given that the data were retrieved from Ctrip.com, this study applies star rating as a primary filtering criterion. In this regard, the list of hotels returned from a customer search can be regarded as a common set of hotels. Within the common set of hotels, customers are presented with several types of important information, such as a brief description of the hotel, lowest price, customer ratings, number of reviews, and customers' recommendation rate. Based on the information provided, customers evaluate the alternatives to form a consideration set—a set of preferred hotels to minimize the risk related to their selection (Mankad et al. 2016; W. Wang, Yi, and Dai 2018). Studies find that customers increasingly rely on online peer-to-peer reviews in their prepurchase evaluation of the hotels within their consideration sets (Filieri et al. 2018; K. Lu and Elwalda 2016). The final decision of a customer is from their consideration set, which derives in turn from the common set of hotels offered by the OTA (Pan, Zhang, and Law 2013).

Methodological Framework

Online review data are now playing an important role in every service industry. It offers an understanding of customer preferences and allows an assessment of a company's reputation (Holloway and Beatty 2003; L. Wu et al. 2016). However, the use of online review data to identify competitors has been overlooked. In this study, we use the hotel industry as an example and present an analytical framework based on a set of machine-learning techniques that will identify a service provider's competitors. It further recognizes the relative importance of different service attributes affecting customers' decision making in various market segments. We demonstrate the applicability of the proposed analytical framework on a sample of over 8 million customer reviews of 6,409 hotels in 50 Chinese cities, taken from Ctrip.com. Given that the service review data are diverse in its format, the underlying analytical

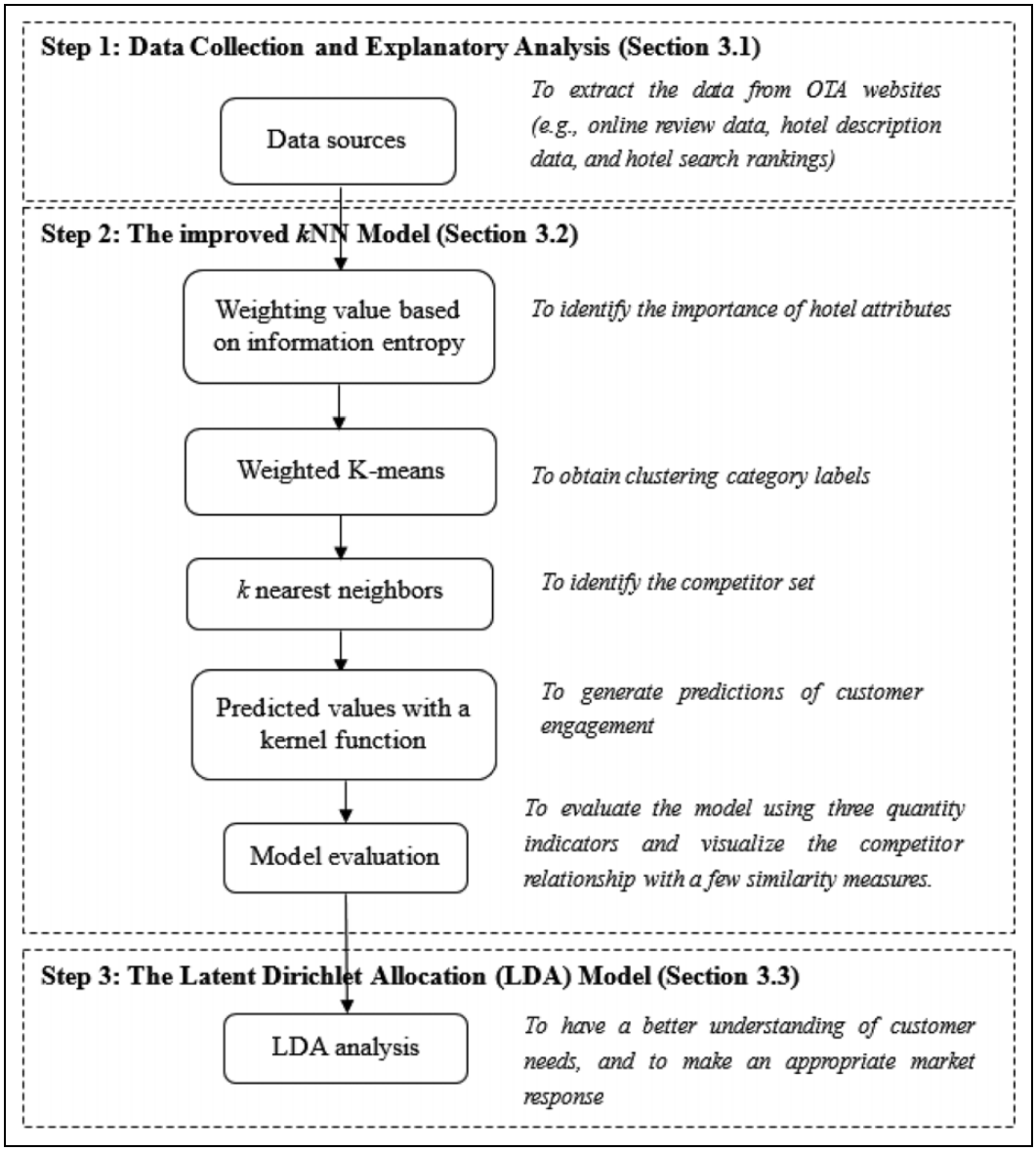

Figure 1. The analytical framework based on the improved *k*-nearest neighbor model and latent Dirichlet allocation model.

framework can help the service provider to identify the competitors in the online battlefield more comprehensively and cost-effectively.

Figure 1 illustrates the overall process of the framework, which comprises three main steps. The first step is to collect all the available data, and Step 1: Data Collection and Exploratory Analysis section explains how both structured data (i.e., customer review ratings) and unstructured data (i.e., customer review text comments) are extracted and used. The second step (Step 2: The Improved *k*NN Model section) is to construct and evaluate an improved *k*NN model for competitor identification by calculating the weighting value to each important attribute of the service based on information entropy to minimize the cross-validation error in the prediction of customer engagement. Based on the competitors identified in Step 2, the third step (Step 3: LDA Model section) uses an LDA model for an in-depth analysis of customer review text comments to get a better understanding of customer needs and competitors' provision to make appropriate market responses.

Step 1: Data Collection and Exploratory Analysis

In this study, we used a data crawler and downloaded all available hotel data from the Ctrip.com website for 50 key tourist cities designated by the China National Tourism Bureau (2016). To ensure the consistency of the data, we consider only the most popular hotels, that is, those on the first 10 pages for each city. It is important to note that the number of hotels listed on each page of Ctrip is fixed at 25 and is not affected by screen size. This gave a total of 12,500 hotels. The hotels with fewer than 100 online reviews and with blank reviews, duplicate hotels, and those lacking complete information (i.e., the

Table 2. Explanation of Structured Variables.

Data Types	Variables	Description
Hotel search ranking	Ranking	Hotel search ranking is the relative position after a search on Ctrip.com
Hotel description data	Star	Two- to five-star hotel rating
	Rooms	To represent the size of the hotel
	Price	The lowest price of a hotel room shown on Ctrip.com
Customer review data	Customer engagement	The number of customer reviews (a proxy measure for online customer engagement) from January to December 2016
	Recommendation	The proportion of all users who have at least reserved the hotel room and would recommend that hotel to others
	Customer rating	The average rating of all reviews of a hotel
	Location convenience	The review rating for the convenience of the hotel's location given by customers
	Staff service	The review rating for the quality of the hotel's staff service given by customers
	Facilities	The review rating for the hotel's facilities given by customers
	Cleanliness	The review rating for the hotel's cleanliness given by customers

number of rooms, price, and recommendation rate) were excluded to improve the validity and reliability of the data. The final data set contains 6,409 hotels with 8,374,102 online reviews posted between January 1, 2016, and December 30, 2016.

As shown in Table 2, for each hotel, we collect three types of structured data, namely, customer review data, hotel description data, and hotel search ranking. Customer review data comprise the number of customer reviews (used as a measure of customer engagement), recommendation rate, the overall customer rating, and a four-dimensional rating of hotel quality (i.e., ratings of the hotel's location convenience, staff service, facilities, and cleanliness). The hotel description data include the hotel's star rating (star), the number of hotel rooms (rooms), and the price of a standard room (price), where a standard room is the primary type offered by each hotel, which is also usually the cheapest. The hotel search ranking (ranking) is included, filtered by Ctrip's hotel star ratings: two-star and below (economy), three-star (comfortable), four-star (high end), and five-star (luxury). Appendix A shows an example of a customer review on Ctrip.com. Reviews can be posted only by users who have at least reserved a hotel on the website, and they all give a summary rating, which can range from one to five. Other than the structured data, we also collect the unstructured textual comments posted by customers to conduct the in-depth text analysis for the identification of key service attributes.

As discussed in Online Reviews for Value Co-Creation and Service Improvement section, customer-company interactions have long been considered an important resource for service providers in value co-creation and service improvement (Brodie et al. 2011). This phenomenon has been further enhanced by the digitization and development of information communication technologies. Traditionally, from the perspective of economic theory, price is treated as a primary strategic variable for hotels, especially in the short term (Weatherford and Bodily 1992), and the intensity of price competition increases when more rooms of similar quality are traded in a relatively small area (Choi 1991). However, online hotel competition today is strongly tied to the OTAs' algorithms, and customer engagement (i.e., the number of reviews) becomes the most important factor for the selected OTA (e.g., Ctrip.com) to consider the hotel's popularity and reflect this in its customer recommendation system where the hotel is ranked in the search return.

Studies show that customer engagement in the online platform reflects the popularity of service providers and can influence as much as 20%–50% of online purchase decisions (Kumar and Pansari 2016; Thakur 2018). Customers are likely to log on to online platforms like Ctrip.com to check reviews as part of their service evaluation (Shao and Kenney 2018; Ye, Law, and Gu 2009). On the one hand, research suggests that online reviews can affect the business of the service industry on a multisided platform like online marketplaces (Gur and Greckhamer 2019; Jin, Ji, and Gu 2016). On the other hand, by posting online reviews, customers can generate important social value within the community (Kumar and Pansari 2016; Thakur 2018). Therefore, this study considers the act of posting online reviews as one of the most influential expressions of customer engagement and takes the total number of customer reviews of each hotel in 2016 as a proxy for online customer engagement. Table 3 displays descriptive statistics for the variables used in the study, and we take customer engagement as the dependent variable in our model.

Step 2: The Improved kNN Model

In order to measure how similar hotels are, we constructed an improved *k*NN model by combining information entropy (Shannon 1948) and weighted *K*-means methods (Modha and Spangler 2003). The *k*NN model is an efficient technique that has been used extensively for classification in machine learning. In essence, when applied for regression, the *k*NN technique makes a prediction based on the *k*NNs in a metric space. However, the standard *k*NN (S-*k*NN) technique uses an exhaustive search of an entire training set and is very sensitive to data sets containing noise (Mitani and Hamamoto 2006) and treats all attributes as having the same importance for prediction results. Although both marketing and service research has recognized the utility of the *k*NN method in analyzing online review data

Table 3. Descriptive Statistics for Variables.

Category	Variable	Min.	Max.	Mean	Standard Deviation	Category	Variable	Min.	Max.	Mean	Standard Deviation
Two-star	Ranking	1	250			Three-star	Ranking	1	250		
	Price	25	653	145.01	58.693		Price	48	2,539	206.04	106.943
	Rooms	5	399	97.26	46.966		Rooms	5	1,100	122.98	75.458
	Recommendation	0.76	1.00	0.9426	0.03635		Recommendation	0.65	1.00	0.9376	0.03871
	Customer rating	3.4	4.9	4.262	0.2310		Customer rating	3.0	5.0	4.229	0.2524
	Location convenience	3.4	4.9	4.329	0.2383		Location convenience	3.1	5.0	4.316	0.2496
	Facilities	3.0	4.9	4.121	0.2846		Facilities	2.7	5.0	4.094	0.3232
	Staff service	3.3	5.0	4.285	0.2431		Staff service	3.1	5.0	4.221	0.2625
	Cleanliness	3.2	4.9	4.313	0.2556		Cleanliness	2.9	5.0	4.281	0.2761
	Customer engagement	88	3,367	914.37	813.133		Customer engagement	67	1,567	696.51	708.372
	N		1,954				*N*		1,580		
Four-star	Ranking	1	250			Five-star	Ranking	1	250		
	Price	170	3,600	306.23	152.740		Price	268	3,105	606.45	335.781
	Rooms	8	1,092	187.84	102.810		Rooms	12	1525	313.98	162.460
	Recommendation	0.74	1.00	0.9521	0.02986		Recommendation	0.17	1.00	0.9686	0.03050
	Customer rating	3.5	4.9	4.314	0.2079		Customer rating	3.8	4.9	4.535	0.1548
	Location convenience	3.3	5.0	4.368	0.2173		Location convenience	3.7	4.9	4.520	0.2020
	Facilities	3.0	4.9	4.200	0.2742		Facilities	3.4	4.9	4.474	0.2043
	Staff service	1.0	5.0	4.298	0.2470		Staff service	3.7	4.9	4.522	0.1586
	Cleanliness	3.5	4.9	4.388	0.2184		Cleanliness	3.8	4.9	4.620	0.1522
	Customer engagement	47	6,607	2,334.08	1,952.063		Customer engagement	145	14,524	3,113.88	2,345.485
	N		1,478				*N*		1,397		

(Arora et al. 2019; Hartmann et al. 2019; Sohn, You, and Lee 2003), the effect of outliers and unimportant attributes are often ignored, which tend to be a critical issue in competitor identification (Baum and Lant 2003).

To enhance the efficiency of competitor identification from a large-scale online review data, an improved *k*NN model was adopted. We apply information entropy to find the relative importance of each focal hotel attribute in the whole data set. Then, we divide the training set into several clusters by the weighted *K*-means clustering method, based on city (the location of the hotels), which helps to overcome the negative impact of outliers in the training set for finding the *k*NNs. That is, only hotels within the same city are considered competitors. Finally, the local mean vector *k*NNs in each cluster are employed to identify competitors of the focal hotel. The improved *k*NN model is constructed as follows.

Given a training data set $T = \{(x_n, y_n)\}_{n=1}^{N}$, where N is the total number of the training data set, $x_n = \{x_n^1, x_n^2, \ldots, x_n^m\} \in \mathcal{R}^m$ is the input hotel with m-dimensional attribute space, and $y_n \in \mathcal{R}$ (function approximation) is the output of customer engagement, and $C = \{C_1, C_2, \ldots, C_i\}$ is the class label for hotels with different star ratings, where i is the number of categories, the clustering category is obtained by the weighted *K*-means method (Modha and Spangler 2003), and $T_i = \{(x_{ij}, y_{ij})\}_{j=1}^{N_i}$ denotes a subset form the class C_i, where $N_i < N$. We search k_i nearest neighbors from the subset T_i for the test hotel x using the weighted Euclidean distance function, defined as follows (Cooper 1983).

$$d(x, x_{ij}) = \sqrt{\sum_{l=1}^{m} w_l (x - x_{ij}^l)^2}, \tag{1}$$

where w_l is the weight ($0 \le w_l \le 1$, $\sum_{l=1}^{m} w_l = 1$) assigned to the lth hotel attribute in different hotel star ratings, representing the relative importance of each hotel attribute. In this study, the well-known information entropy is applied to evaluate the attribute weight w_l of different hotel star ratings, which minimizes the sum of square error in the process of constructing the improved *k*NN model. The information entropy method is frequently used to determine the weights of different objectives in decision-making models. The weights of hotel attributes based on the information entropy method in different hotel star ratings are defined in Appendix B.

Sorting the distances of each subset in ascending order, we can get the nearest neighbor set for the test hotel x in the subset T_i marked as $T_{ik}^{NN} = \{(x_{ij}^{NN}, y_{ij}^{NN})\}_{j=1}^{k_i}$. We then utilize the kernel function to estimate the predicted value for the subset T_i based on its neighbors, which can solve the multicollinearity between the input variables. For the test hotel x, a Gaussian radial basis function as the kernel function is expressed as: $K(x, x_{ij}^{NN}) = \exp(-||x - x_{ij}^{NN}||^2/2\sigma^2)$, where σ is a smoothing parameter. If the value of σ is appropriately selected, it can compensate for k_i exceeding the permitted

value. In the experiments, σ is set to be half the mean distance between x and k_i nearest neighbors of each subset. The prediction for testing hotel x, described as $\hat{y}_i$ in the subset T_i, is defined as follows:

$$\hat{y}_i = \frac{\sum_{j=1}^{k_i} y_{ij}^{NN} K(x, x_{ij}^{NN})}{\sum_{j=1}^{k_i} K(x, x_{ij}^{NN})}. \quad (2)$$

In the process of calculating the predicted value $\hat{y}_i$, the attribute weight w_l and the number of nearest neighbors k_i are the important parameters, as they control the flexibility of the improved *k*NN model. The value of w_l is determined by the information entropy method as above. The value of k_i determines the accuracy and smoothness of the predicted values in the subset T_i, which are calculated by 10-fold cross-validation (Golub and Wahba 1979).

After getting the k_i value in the subset T_i, we calculate the local mean vector $U_{ik} = \frac{1}{k_i}\sum_{j=1}^{k_i} x_{ij}^{NN}$ in the subset T_i. Finally, we assign the *kNN*s of the test hotel x by calculating the closest distance to the local mean vector of U_{ik}. Thus, we obtain k competitors of the test hotel x.

$$k = \arg\min_{k_i} d(x,\ U_{ik}). \quad (3)$$

The key advantages of the improved *k*NN technique are the comprehensive utilization of information entropy for calculating the weight of each hotel attribute and the use of the local mean method for obtaining the nearest neighbor k not from all training data but from each cluster of the training data.

This study applies two approaches to evaluate the reliability and validity of the improved *k*NN model for analyzing the competitor sets of focal hotels. First, we compare the improved *k*NN model with the S-*k*NN model (X. Wu et al. 2008) and the competitive linear regression (LR) model (Ritov 1990). The correlation coefficient (CC), mean absolute error (MAE), and root mean square error (RMSE) are the three main indicators for the comparison. Specifically, CC measures the degree of correlation between predicted and observed values, while MAE and RMSE measure the deviation in the observed and predicted values. The formulas for the three indicators are provided in Appendix C. The model with the highest value for CC and the lowest values for MAE and RMSE is considered the best.

Then, this study weights each hotel attribute differently to reduce the prediction bias of the improved *k*NN model. This is because customer behavior regarding each hotel attribute could differ across market segments. We further compared the applicability of our improved *k*NN model with different commonly used similarity measures such as Euclidean distance, cosine similarity, and Pearson CCs. This is important, as the application of the similarity measurement method is closely related to the data analysis process and therefore can significantly affect the outcomes. To justify the assertion that the improved *k*NN model is the most appropriate approach, we conducted similarity and divergence analysis on the comparison between the competitive map generated by our improved *k*NN model and the MDS perceptual maps generated by the aforementioned similarity measures.

Step 3: LDA Model

The process of competitor identification uses the quantitative hotel attribute information from the customer reviews and then combines the hotel description data and search ranking information to analyze the competitive relationships among hotels. Although such an analytical method provides managers with an effective way to scan the market for competitors, it overlooks the textual information in customer reviews (i.e., the unstructured data) and in particular the rich information on hotel attributes. According to Mankad et al. (2016), the textual content of customer reviews has more customer insight than the quantitative hotel attribute rating. Therefore, we take the unstructured data in customer reviews as the object and use the LDA model to extract customer insights from the textual comments. The LDA model is a powerful and widely used topic-modeling algorithm (Blei 2012; Blei, Ng, and Jordan 2003). It constructs a three-layer Bayesian structure of documents, topics, and key words and regards documents as a probability distribution of implicit topics and topics as a probability distribution of key words. In addition, the distribution of key words for different topics varies, so all reviews can be viewed as consisting of two probability distributions: $p(\text{word}|\text{review}) = \sum_{\text{topic}} p(\text{word}|\text{topic}) \times p(\text{topic}|\text{review})$.

For a given set of text data, LDA uses a probabilistic framework to infer the set of hidden topics from the customer reviews and decomposes each review into a mixture of these topics with different probabilities (Blei 2012; Blei, Ng, and Jordan 2003). The text information from customer reviews is clustered into different topics, and the attribute key word vector of the data is constructed from the topics. The focal hotel can then benchmark its service against that of competitors and pay more attention to the relevant topics.

Results

We conduct an in-depth analysis of the data (including hotel description data, hotel search rankings, and review comments) and empirically test how well the analytical framework can identify key competitors and determine the importance of particular hotel attributes in different market segments. The outputs can be used to identify the focal hotel's strengths and weaknesses, with visual representation of the results, all of which support more informed strategic marketing decisions. Moreover, this study also takes an unstructured view to harvest customer comments from the competitors' online reviews. It allows managers to identify "hot topics" that capture users' perceptions of the hotel and compare the "hot topics" among competitor hotels to make an appropriate market response within specific market segments.

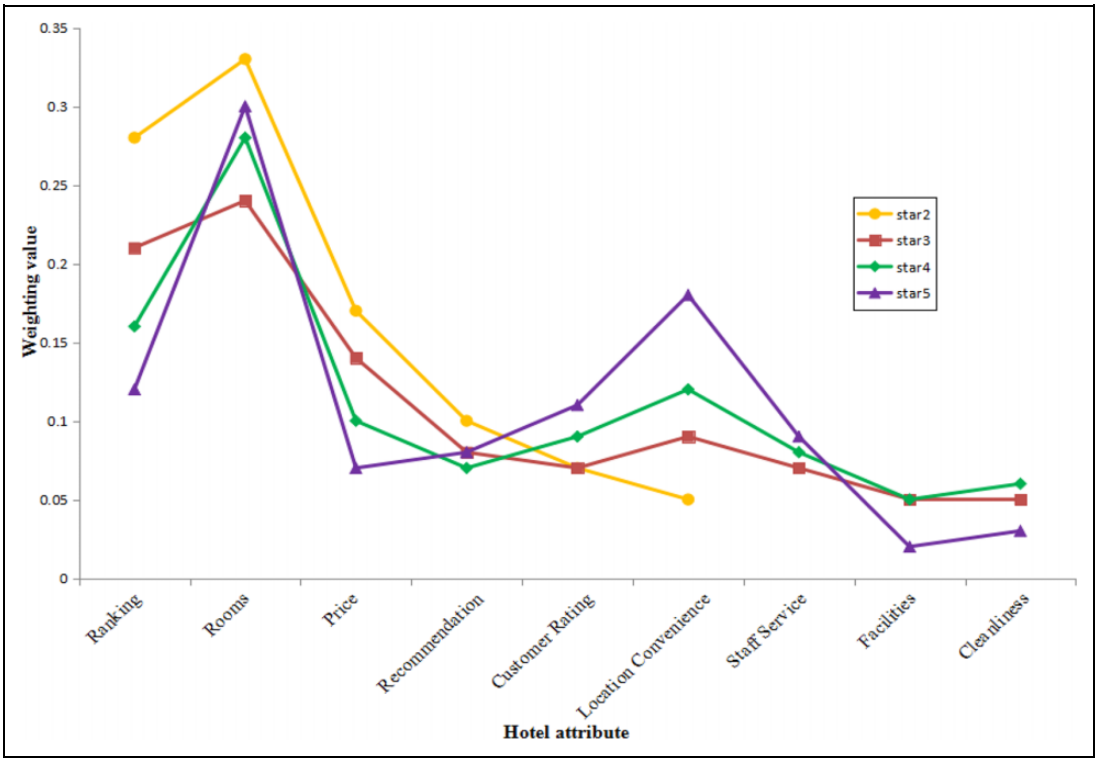

Figure 2. The importance of hotel attributes for different hotel star ratings.

Identifying the Importance of Hotel Attributes

After the data collection, all variables are normalized to prevent those with a high variance from dominating those with a lower one. The first step of the improved *k*NN model is to calculate the information entropy of attributes and get their corresponding weights in different market segments (according to the star rating of hotels). Figure 2 shows the weight of each hotel attribute for different hotel star ratings.

From the results, rooms is the most important attribute for all hotels, from two-star to five-star, with weight values of 0.33, 0.24, 0.28, and 0.30, respectively. In terms of other attributes, the outcomes are quite different for hotels with different star ratings. For two-star hotels, ranking (0.28) is the second most important attribute and has a more important influence on customer engagement than for other hotels, which is also true of price (0.17). Compared with two-star hotels, the weight of ranking (0.21) for three-star hotels is reduced but is still second in importance. Price (0.14) is third, followed by location convenience (0.09), recommendation (0.08), staff service (0.07), and customer rating (0.07). For four-star hotels, the weights of ranking (0.16) and price (0.10) are lower than for two- and three-star hotels. They both become less important in this segment. Additionally, the weight of location convenience (0.12) is higher, and location becomes the third most important attribute, followed by a cluster of three attributes, recommendation (0.07), customer rating (0.08), and staff service (0.08). For five-star hotels, location convenience (0.18) is the second most important attribute and is higher than for hotels with other star ratings. Note that ranking (0.12) and price (0.07) are less important for five-star hotels than they are for other hotels. Nevertheless, ranking is still the third most important attribute for five-star hotels. It is followed, in order, by recommendation (0.11), customer rating (0.08), and staff service (0.09), and the remaining attributes consist of cleanliness (0.03) and facilities (0.02).

Identifying the Competitor Set

The second step is to divide the training data set into four clusters based on the weighted *K*-means method and to construct the improved *k*NN model with weighted attributes in each subset in order to identify the *k* competitors of the focal hotel. The value of the parameter *k* is selected to give the model with the smallest sum of squares error in the cross-validation. For two-star through to five-star hotels, the *k* values selected are 8, 6, 16, and 9, respectively. In other words, from the perspective of the customer, a focal hotel's managers in different market segments can confirm exactly how many key competitors it is likely to have: as few as six for a three-star hotel and as many as 16 for a four-star hotel. However, the optimal value of *k* may differ from one data set to another (D. Cheng et al. 2014).

Generally, when a customer inputs the name of a focal hotel in the search box on Ctrip.com, other hotels will additionally appear among the search results, as shown in Appendix D. These additional hotels are recommended according to the search algorithm of Ctrip.com. Moreover, customers tend to include the top-ranked hotels recommended by an OTA in their

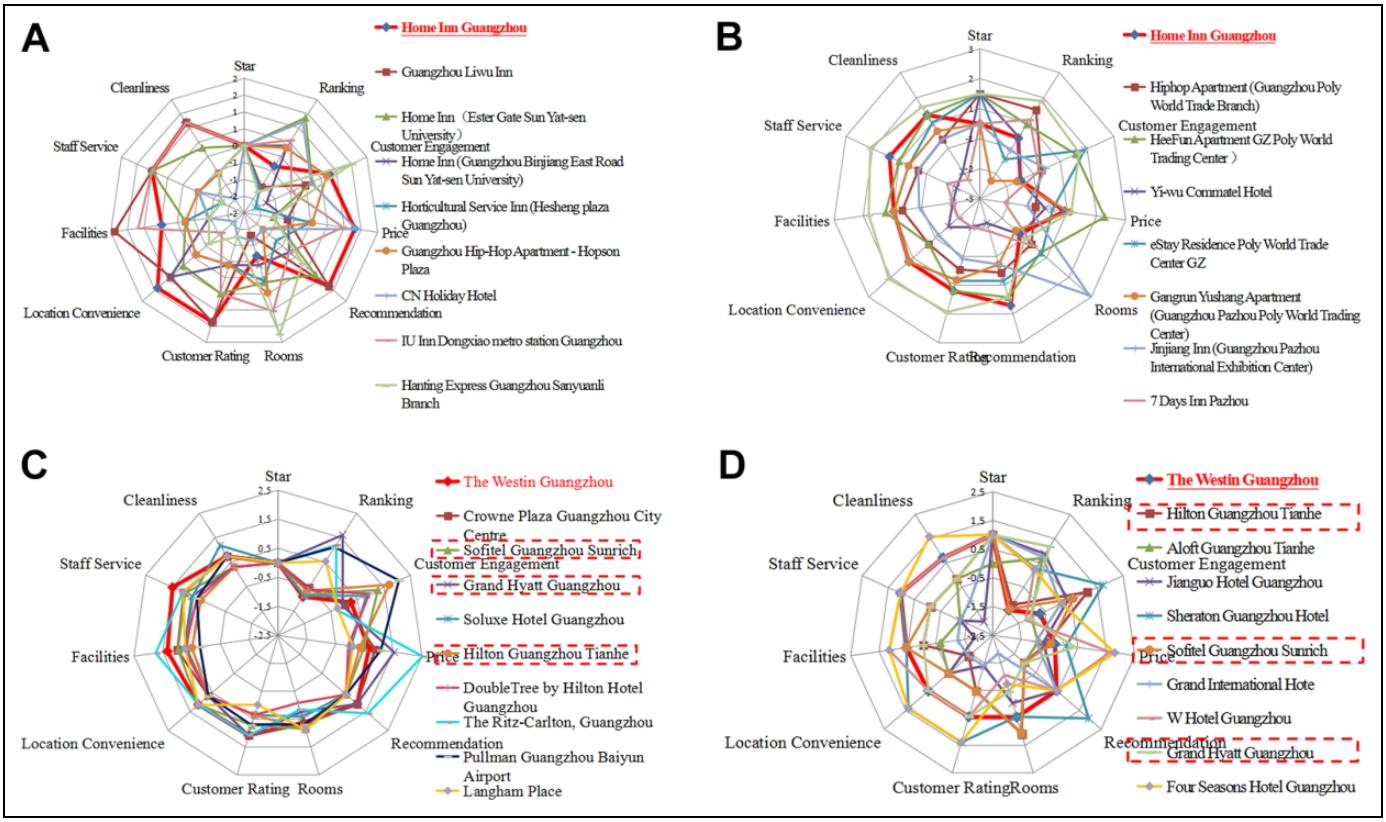

Figure 3. Identifying the competitor set with different attributes. (A) Competitor sets of Home Inn Guangzhou based on the improved *k*NN model. (B) Popular nearby hotels of Home Inn Guangzhou recommended by the OTA. (C) Competitor sets of the Westin Guangzhou based on the improved *k*NN model. (D) Popular nearby hotels of the Westin Guangzhou recommended by the OTA. *Note*. *k*NN = *k*-nearest neighbor; OTA = online travel agent.

consideration sets for a final selection decision (Chen and Yao 2016). Consequently, it is of interest to examine both customers' and the OTA's competitor sets (i.e., competitors identified from the two different perspectives) and to compare "matches" and "mismatches." The purpose of this comparison is to evaluate the effectiveness of the managerial competitor identification model proposed in this study (based on the customer perspective) and to find the reason for any "mismatch."

We randomly take a two-star hotel, Home Inn Guangzhou, and a five-star hotel, the Westin Guangzhou, as two focal hotels, and identify their competitor sets using the proposed model and the OTA recommendations. In Figure 3, focal hotels are highlighted in red, and the hotels within red dotted boxes are competitors that are identified (matched) from both the *k*NN model and the OTA recommendations. For instance, Figure 3A and B shows the competitor sets of Home Inn Guangzhou, identified by the improved *k*NN model and by the OTA, respectively. We compare two lists of competitors and find there is no hotel in common. Nonetheless, Figure 3C shows the list of competitors of the Westin Guangzhou identified by the improved *k*NN model, in which there are three hotels consistent with the list of hotels by recommended the OTA, as shown in Figure 3D.

In addition, we note that Home Inn Guangzhou has advantages on the customer rating, location convenience, staff service, cleanliness, and recommendation attributes, and the room price is at a medium to high level in Figure 3A. Although Home Inn Guangzhou has these advantages compared with its competitors, it is not the most popular hotel in the market. A practical recommendation for the hotel managers of Home Inn Guangzhou would be to prioritize its investment in its search ranking and facilities to achieve more customer engagement. It also can be noted that the popular nearby hotel sets recommended by the OTA include three-star as well as two-star hotels. Thus, we find that the OTA has a tendency to cross-sell other star-rated hotels to customers. In Figure 3C, our approach identifies that staff service is the Westin Guangzhou's distinct advantage, and the location and customer rating attributes are superior, but a few competitors have similar characteristics to the focal hotel. The facilities and recommendation attributes and customer engagement are at a medium to high level, while the rest are at or below the average level of competitors. We conclude that the Westin Guangzhou has better customer engagement than its competitors because of its competitive advantages in staff service, location convenience, and customer rating but needs to address weaknesses concerning online search ranking, room cleanliness, and pricing.

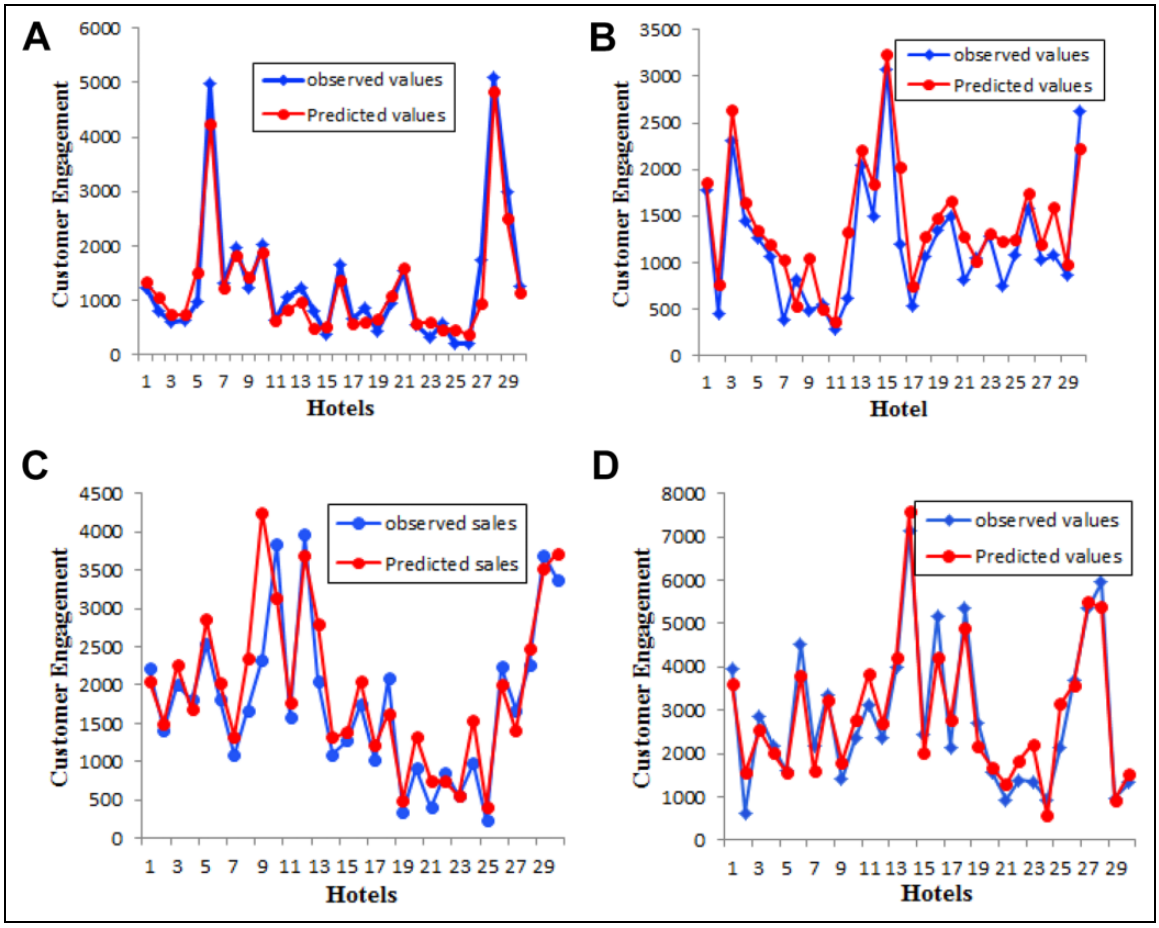

Figure 4. Predicting hotel customer engagement for different hotel star ratings. (A) Two-star hotels. (B) Three-star hotels. (C) Four-star hotels. (D) Five-star hotels.

Predicted Customer Engagement

After k and w_l are obtained, the third step is to predict customer engagement based on the improved kNN model. Figure 4 illustrates the prediction results for 30 hotels in each of the four categories in the test data set. It can be seen that the improved kNN model closely tracks the trend in customer engagement for all 120 hotels. This further confirms the ability of our model to identify competitors across different hotel segments.

Model Evaluation

We evaluate the performance of the improved kNN model against two conventional benchmark models (i.e., LR and S-kNN) in its prediction capability. Table 4 presents the results of three indicators used for all models. We can see that all the CC values of the improved kNN model are more than 0.6, which indicates there are moderate to strong linear correlations between the predicted and the observed data. Indeed, the CC values of the improved kNN model are higher than those in LR and S-kNN models for hotels of all star ratings, indicating that the improved kNN model outperforms the other two models in terms of the CC indicator. In the evaluation of the MAE and RMSE indicators, compared with the benchmark models, the improvements of the proposed model on average are 49.51% and 32.63% for the LR model and 38.58% and 26.46% for the S-kNN model, respectively, for all hotels of any star rating. For five-star hotels, the indicators are improved by 59.11% and 48.36% for the MAE and 44.40% and 40.18% for the RMSE, using the improved kNN model compared with the LR and S-kNN models, respectively. Overall, the improved kNN model achieves the highest values on the CC indicator and the lowest values on the MAE and RMSE indicators and performs significantly better than the benchmark models in its predictions.

To benchmark our improved kNN model, we compared the weighted Euclidean distance similarity measure applied in our model with other measures (i.e., standard Euclidean distance, cosine similarity, and Pearson CC) that are commonly used in conventional MDS techniques to capture the similarity relationships among products. We use Guangzhou Westin Hotel as the focal hotel and randomly identify 101 five-star hotels in Guangzhou to demonstrate the results obtained using different similarity measures. The results are shown in Appendix E. Specifically, Figure E1A was generated based on the weighted Euclidean distance while Figures E1B–D show the MDS perceptual maps generated through standard Euclidean distance, cosine similarity, and Pearson CCs. Figure E1A uses price and

Table 4. Comparison of the Improved *k*NN With Two Benchmark Models.

Indicator	Linear Regression	S-*k*NN	Improved *k*NN
	Two-star		
CC	.672	.659	.783
MAE	.356	.294	.180
RMSE	.530	.430	.361
	Three-star		
CC	.575	.615	.681
MAE	.470	.296	.258
RMSE	.720	.542	.437
	Four-star		
CC	.613	.645	.717
MAE	.564	.425	.314
RMSE	.648	.569	.397
	Five-star		
CC	.595	.632	.859
MAE	.384	.304	.157
RMSE	.455	.423	.253

Note. *k*NN = *k*-nearest neighbor; S-*k*NN = standard *k*-nearest neighbor; CC = correlation coefficient; MAE = mean absolute error; RMSE = root mean square error.

location as two hotel attributes and classifies the hotels into four different clusters on an XY graph. The focal hotel is shown as a green triangle (Cluster 3), while the red bubble was generated by the weighted Euclidean distance to capture the nearest-neighbor competitive hotels. The focal hotel's cluster is positioned in the right upper quadrant, which indicates that it was classified as having a high recommendation rating (i.e., 4.7 of 5) toward the hotel location and a relatively high average price (i.e., 1,121 RMB). However, given the same level of data attributes and sample size (which is relatively small in this example), the perceptual maps generated in Figures E1B–D tend to be difficult to interpret.

Robustness Check

After evaluating the model, we perform a check to ascertain that the stability of the improved *k*NN model is robust to data sets from two different time periods. To this end, we harvest online customer reviews, hotel description information, and hotel searching rankings from January to December 2018 on Ctrip.com, for comparison with the results from the 2016 data set reported above. These two overall data sets, though, are not wholly comparable, because, for example, in the intervening period (2017), some hotels may have closed or ceased to list themselves on Ctrip.com. Therefore, 50 of the hotels that appeared in both data sets were randomly selected for each of the four market segments (hotel star ratings). We use the analysis of variance test to compare the two error groups across the four market segments. As shown in Table 5, all the p values are above .05, thus indicating there is no significant difference in error value between the two data sets.

As a further check on robustness, we pick 30 hotels at random and plot the 2016 and 2018 predicted values in Appendix F. The plots are similar, which again indicates that the improved *k*NN model is robust, and the results it generates apply equally well to the Chinese hotel competitive environment in 2018 as to that in 2016.

Table 5. One-Way Analysis of Variance of 2016 and 2018 Results for the Different Hotel Star Rating.

	Source of Variation	Sum of Squares	*df*	Mean square	*F*	*p* Value
Two-star hotels	Between groups	22.160	1	22.160	2.325	.758
	Within groups	934.055	98	9.531		
	Total	1,660.720	99			
Three-star hotels	Between groups	98.240	1	98.240	3.230	.213
	Within groups	7,827.250	98	30.415		
	Total	7,925.490	99			
Four-star hotels	Between groups	36.572	1	36.572	2.893	.652
	Within groups	1,238.871	98	12.641		
	Total	1,275.443	99			
Five-star hotels	Between groups	19.827	1	19.827	1.678	.893
	Within groups	1,157.953	98	11.816		
	Total	1,177.780	99			

Unpacking Customer Reviews

We used a data crawler and downloaded all customer reviews of the Westin Guangzhou, and its nine competitors identified from the improved *k*NN model, on Ctrip from January 1, 2016, to December 31, 2016. In total, the data sample comprised 14,897 online reviews across these 10 hotels.

The "perplexity" value was used as the benchmark to determine the number of topics (Blei, Ng, and Jordan 2003). The smaller the perplexity value, the better is the fitness of the model with different numbers of topics. Consistent with the study of Hoffman, Bach, and Blei (2010), the perplexity value was evaluated using five-fold cross-validation, and the results suggest the five most appropriate topics for the LDA model used in this study. The five topics were interpreted as location, amenities, value, experience, and transaction. Specifically, location is the place where the hotel is situated and whether it is convenient for customers; amenities indicates the useful services and features provided when staying at the hotel; value is associated with the customer perceived value for money after or during the hotel stay; experience mainly refers to the overall experience of the customer's stay; transaction is mostly about transactional behaviors and the mechanics of the customer's stay (it mostly appears during check-in or check-out and/or before customers arrive at the hotel). These topics capture most of the textual information in customer reviews. However, these

Table 6. Most Likely Words in Each Topic.

Topic1 (Location)	Topic 2 (Amenities)	Topic 3 (Value)	Topic 4 (Experience)	Topic 5 (Transaction)
Location	Facilities	Price	Room	Front
Traffic	Breakfast	Cost	Service	Check
Near	Floor	Place	Staff	Night
Convenience	Bathroom	Economy	Clean	Room
Metro	Enthusiasm	Cheap	Comfortable	Speed
Station	Swim	Star	Attitude	Upgrade
Restaurant	Bed	Recommend	Quiet	Service
Supermarket	Wi-Fi	Free	Every	Call
Walk	Large	Old	Considerate	Taxi
Scenery	Surroundings	Satisfaction	Nice	Lobby

topics are destination specific, in that they may not be the same if our data had been collected from a different set of hotels, particularly for location (Topic 1) and experience (Topic 4). Each topic contains different attribute key words with particular probabilities that the key words belong to that topic. Table 6 lists the 10 attribute key words that were most likely to appear in each topic, in descending order.

The strength of each topic can be computed by the LDA model, and Table 7 compares topic strengths for the Westin Guangzhou with those of its nine competitors identified from the improved *k*NN model. A larger value of the topic strength represents a more popular (more often discussed) topic. In this study, the strength of the topic for each hotel is closely related to the popularity of the topic in the 2016 time period. In other words, the topic strength in 2016 is proportional to the number of reviews discussing or at least mentioning that topic.

Based on the topic strengths of different hotels, we conducted additional analysis to classify and examine the sentiments of the customer reviews to further understand each topic identified. In this way, we used the overall hotel ratings as a proxy for the "emotion recognition" of those review comments (Liu 2006; Ye, Law, and Gu 2009) and classified comments on hotels with a star rating above the overall average online review score as positive, while review comments on hotels below a star rating of three were assumed to be negative comments; the rest were designated neutral comments. As neutral comments presumably do not significantly drive customer behavior (Liu 2006), this study considers only positive and negative comments to analyze the preferences of customers. Appendix G shows two examples (the Westin Guangzhou Hotel and its major competitor, the Sofitel Guangzhou Sunrich Hotel) of co-occurrence networks generated using this approach. The size of the node represents the frequency of the key words, and the line thickness represents how often particular pairs of key words occurred in the same comment.

Table 7. The Strengths of the Five Topics Across 10 Competing Hotels.

	Location	Amenities	Experience	Value	Transaction
The Westin Guangzhou	.0461	.0565	.0612	.0236	.0198
Crown Plaza Guangzhou City Centre	.0467	.0256	.0527	.0345	.0223
Sofitel Guangzhou Sunrich	.0689	.0193	.0458	.0476	.0341
Grand Hyatt Guangzhou	.0223	.0547	.0578	.0178	.0289
Soluxe Hotel Guangzhou	.0543	.0376	.0258	.0298	.0241
Hilton Guangzhou Tianhe	.0345	.0569	.0398	.0352	.0167
Double Tree by Hilton Hotel Guangzhou	.0312	.0329	.0213	.0467	.0201
The Ritz-Carlton, Guangzhou	.0421	.0543	.0644	.0246	.0312
Pullman Guangzhou Baiyun Airport	.0568	.0513	.0467	.0294	.0245
Langham Place	.4370	.0317	.0289	.0214	.0218

Discussion and Implications

By analyzing over 8 million customer reviews extracted from Ctrip.com, one of the world's largest hotel OTAs, the calculated weights for different attributes reveal customers' preferences regarding hotel selection. "Rooms" is the most important attribute affecting customer engagement across all star ratings. The number of hotel rooms is closely related to the star rating, in that four- and five-star hotels have significantly more rooms than two- and three-star hotels. This indicates that hotels with more rooms are likely to receive more reviews, which in turn leads to more bookings being made. This finding concurs with that of Phillips et al. (2015). It is important to note that once hotels are built, the rooms attribute is fixed. However, it is not reasonable to ignore this attribute, especially for a hotel chain, as managers should consider it in their site selection and procurement of hotels (Song and Ko 2017).

The findings show that the importance of particular hotel attributes varies across different hotel segments according to hotel star ratings. For two-star hotels, ranking is the second most important attribute after rooms. As two-star hotels account for a large proportion of the total number of city hotels, their higher ranking in the search results and higher number of hits per hotel make it easier for them to be included in customers' consideration sets (Chen and Yao 2016). Therefore, the managers of budget hotels should have a strategy to optimize their return in a search through an OTA. Also, the results indicate that in this market segment, price is of concern to customers, more so than location, but this finding is the opposite of

that reported by Mohammed, Guillet, and Law (2014). Similar to two-star hotels, the ranking attribute is important for three-star hotels. The customers are likely to have greater expectations of a three-star hotel, for instance, in terms of staff service, facilities, and cleanliness. Regarding the customers of four-star hotels, hotel location (close to scenic resorts, a commercial district, and a public transport hub) becomes more important than the price when they select a hotel. Also, customers pay more attention to recommendation, customer rating, and staff service.

For five-star hotels, although location is only the third most important attribute, it has a higher weight than for hotels of other star ratings. The price and ranking attributes are given lower weights than for other hotels. On the one hand, this suggests that customers are willing to spend more on a convenient location. On the other hand, Pavlou and Dimoka (2006) point out that the information on these luxury hotels can be easily found on OTAs' websites, as there are a relatively small number of them in a given city. We also note that the five-star hotels have a lower weight on staff service, facilities, and cleanliness than on location convenience, which is counter to the findings of Nasution and Mavondo (2008). A possible explanation is that the customers take it for granted that a luxury hotel will offer high-quality services and facilities, and so they pay more attention to hotel location instead.

Apart from important attributes identified among different hotel segments, this study compared the list of competitors identified from a customer perspective with the list of the hotels recommended by the OTA (i.e., Ctrip.com) and found that the two lists of hotels do not completely match. Thus, there may be inconsistencies between the consumer perspective and OTAs' interests. Additionally, although the proposed *k*NN method provides managers with a way to quickly scan their market competition, there could be a limitation on predictions of individual behaviors or perceptions. Thus, to obtain a more in-depth knowledge of their customers, we proposed a natural-language processing technique (i.e., the LDA model) to analyze online customer reviews. In the case demonstration of the Westin Guangzhou hotel, using the LDA model, we can compare the topic strength of the Westin Guangzhou with its nine competitors identified from the *k*NN model. Based on the LDA analysis, we further compared the co-occurrence network maps between the Westin Guangzhou Hotel and its major competitor, the Sofitel Guangzhou Sunrich Hotel (in Appendix G). The results indicate that the Westin Guangzhou Hotel performed better than the Sofitel Guangzhou Sunrich Hotel on the topic of amenities, while it performed less competitive on the topic of location. Also, the transaction-related key words indicate that neither hotel does well on this topic. This, in turn, means if one of the hotels can improve its transaction quality, such as increasing the speed at reception, this may help it to gain significant competitive advantage over the other hotel.

Implications for Research

This study extends the application of online reviews in service research. Although other recent studies have given attention to competition analysis, service research has mostly involved analysis of survey and archival data (Wieringa and Verhoef 2007; J. Wu and Olk 2014), and the findings are limited by the sources of data (such as cross-sectional data and small sample sizes) and simplistic approaches to the analysis (such as the use of ordinary least squares regression models). As a result, the conclusions usually are neither reliable nor robust (Gao et al. 2018). In contrast, this study extends the literature by proposing and verifying an analytical framework based on a set of machine-learning techniques that has rarely been applied in the competitive environment of service industries (Gur and Greckhamer 2019). We proposed an improved *k*NN model that captures the complex dependency of customer review data, hotel search rankings, and hotel descriptions to visualize the advantages and weaknesses of consumers' perceived service performance and identify key competitors in the marketplace. The results were further applied in a natural-language processing method—the LDA model—to identify the key service topics discussed in the customer reviews of competitors' services. Given that the service review data are diverse in its format, the underlying analytical framework outperforms the other typical machine-learning models and can help the service provider to identify the competitors in the online battlefield more comprehensively and cost-effectively.

Moreover, the view that approximates the customer perspective on competitors can reduce managerial "blind spots," shortsightedness, and competition asymmetry (Baum and Lant 2003), which is consistent with the study of Li and Netessine (2012). While previous studies have identified that attributes such as location, company size, price, and service are often the main factors that define competitors in the hotel industry (J. Y. Kim and Canina 2011), this study indicates that the importance of these attributes varies with hotel star ratings. Last but not least, according to the analytical framework, the market environment can be displayed graphically. The proposed approach can also be adopted in different service industries to determine the perceived quality of services and develop an effective strategy for service improvement.

Implications for Practice

Our findings have important implications for service managers, online consumers, and OTAs in harvesting online reviews to improve their service performance and decision making.

First of all, the proposed analytical framework can help service managers to gain a better understanding of their key competitors as well as customers to make appropriate market responses in a timely manner. Online reviews contain vast amounts of information and reflect customers' demand preferences for hotels. The improved *k*NN model helps managers to determine the degree of influence of particular hotel attributes on consumer decisions across different hotel star ratings. Combined with the hotel's customer engagement, managers can explore the attributes of their key competitors that customers discussed in their reviews. Moreover, the LDA model enables the managers to cluster online reviews into different topics

based on popularity and the sentiments expressed by customers. In this way, tracking changes in the overall content of customer reviews can help managers develop competitive business and management strategies. Furthermore, the co-occurrence analysis can help managers to determine major issues, so that they can pay particular attention to them. For instance, managers can apply the approach to monitor topics with negative sentiment and take action to address the negative comments to minimize adverse outcomes. Managers can also monitor the reviews over time to find out whether their actions generate actual business improvements. Thus, managers can clarify the hotel's position in the market against competitors and fully understand their own advantages and weaknesses, as well as the attitudes of customers, which can guide hotels in service improvement.

The developed analytical framework also provides clear managerial implications for online consumers and OTAs by not only illustrating an effective approach but also producing several visual analytics as examples to follow. In particular, the developed analytical framework consists of competitor set graphs (as shown in Figure 3), LDA topic modeling (presented in Table 6) together with the key word co-occurrence networks (in Appendix G). These could be developed into a software application. Such an app would help online consumers to conduct content analysis and offer valuable insights by harvesting a large number of reviews from different platforms to support their hotel selection. Furthermore, the analytical framework developed in this study can be applied within a broad range of fields to support OTAs' information management, processing, and interpretation. For example, the analytical approach can be adopted by OTAs to improve their analytics. In this way, it can further enhance their operational practices to offer better search results.

Conclusions and Future Research Directions

The purpose of this study is to provide insights into the procedures that could be used by service managers to identify competitors and recognize the relative importance of different product attributes based on online customer reviews. We propose an analytical framework based on a set of machine-learning techniques, including an improved *k*NN model and an LDA model, to identify both the competitor set and important attributes in different star-rated segments of the hotel market. The information on important attributes, each with its own weight, makes this an innovative approach to the identification of key competitors from the perspective of customers (Li and Netessine 2012). We also tested the prediction accuracy, reliability, and robustness of our proposed method. We further use the LDA model for an in-depth analysis of customer review text comments to identify the key topics discussed in competitors' reviews and to make appropriate market responses for improved service competitiveness. While the proposed analytical framework is potentially useful, there are a number of research issues that remain to be addressed.

First of all, this study identifies competitors from the customer perspective. However, as competitors can be determined from different perspectives, it might be difficult for managers to agree with the results generated from our proposed framework. Thus, further research should include other perspectives (e.g., the managers' perspective) to identify other competitor sets, and then compare these, so as to formulate more accurate marketing strategies. Secondly, this study could be extended by segmenting the market in ways other than star ratings, for example, by hotel brand or type (e.g., business and leisure). In this way, future research could more systematically analyze competitors from multiple perspectives for multiple types of market segmentation. Finally, we use hotel data collected from Ctrip.com, which is one of the largest OTAs in China. But taking data from just one source may make the results prone to bias. Future studies can be conducted to verify the analytical framework using data from different sources. Given the dramatic rise of peer-to-peer services (e.g., Airbnb and Flipkey.com) in the hospitality marketplace, the developed framework should be used to investigate the impact of the sharing economy by comparing hotels at a specific destination, available via OTAs (e.g., Expedia and Ctrip.com), with accommodation available through peer-to-peer platforms. This would be of particular interest, as studies indicate that peer-to-peer (sharing) platforms offer a broader range of products and services than traditional OTAs (M. Cheng 2016). Our proposed analytical framework could be used to study the sharing economy and determine the potential changes to the customer experience through the use of different peer-to-peer services.

Appendix A

Figure A1. An example a customer review.

Appendix B

Attribute Weights w_l

The information entropy of hotel attributes for each star-rated hotel data set is defined as:

$$H_l = -\sum_{j=1}^{n} p_j^l \log_2^{p_j^l},\ j = 1, \ldots, n; l = 1, 2, \ldots, m,$$

where $p_j^l = x_j^l / \sum_{j=1}^{n} x_j^l$, and then the hotel attribute weight can be expressed as:

$$w_l = \frac{1 - H_l}{m - \sum_{l=1}^{m} H_l},\ l = 1, 2, \ldots, m,$$

where $0 \leq w_l \leq 1$, $\sum_{l=1}^{m} w_l = 1$.

Appendix C

Three Indicators for Model Evaluation

Given a pair of random variables $(y_i, \hat{y}_i)$, the formula for the CC is:

$$\mathrm{CC} = \frac{\mathrm{Cov}(y_i, \hat{y}_i)}{\sqrt{\mathrm{Var}[y_i]\mathrm{Var}[\hat{y}_i]}},$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, Cov is the covariance, $\mathrm{Var}[y_i]$ is the variance of y_i, and $\mathrm{Var}[\hat{y}_i]$ is the variance of $\hat{y}_i$. MAE is a measure of the difference between two continuous variables. Assume y_i and $\hat{y}_i$ are the variables of paired observations that express the same phenomenon. The MAE is given by:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|.$$

RMSE is a frequently applied measure of the difference between values predicted by a model and the values observed. For example, the RMSE of predicted values y_i for i instances of a regression's dependent variable, $\hat{y}_i$ with variables observed over N times, is:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}.$$

Appendix D

Figure D1. An example of search results from Ctrip.com.

Appendix E

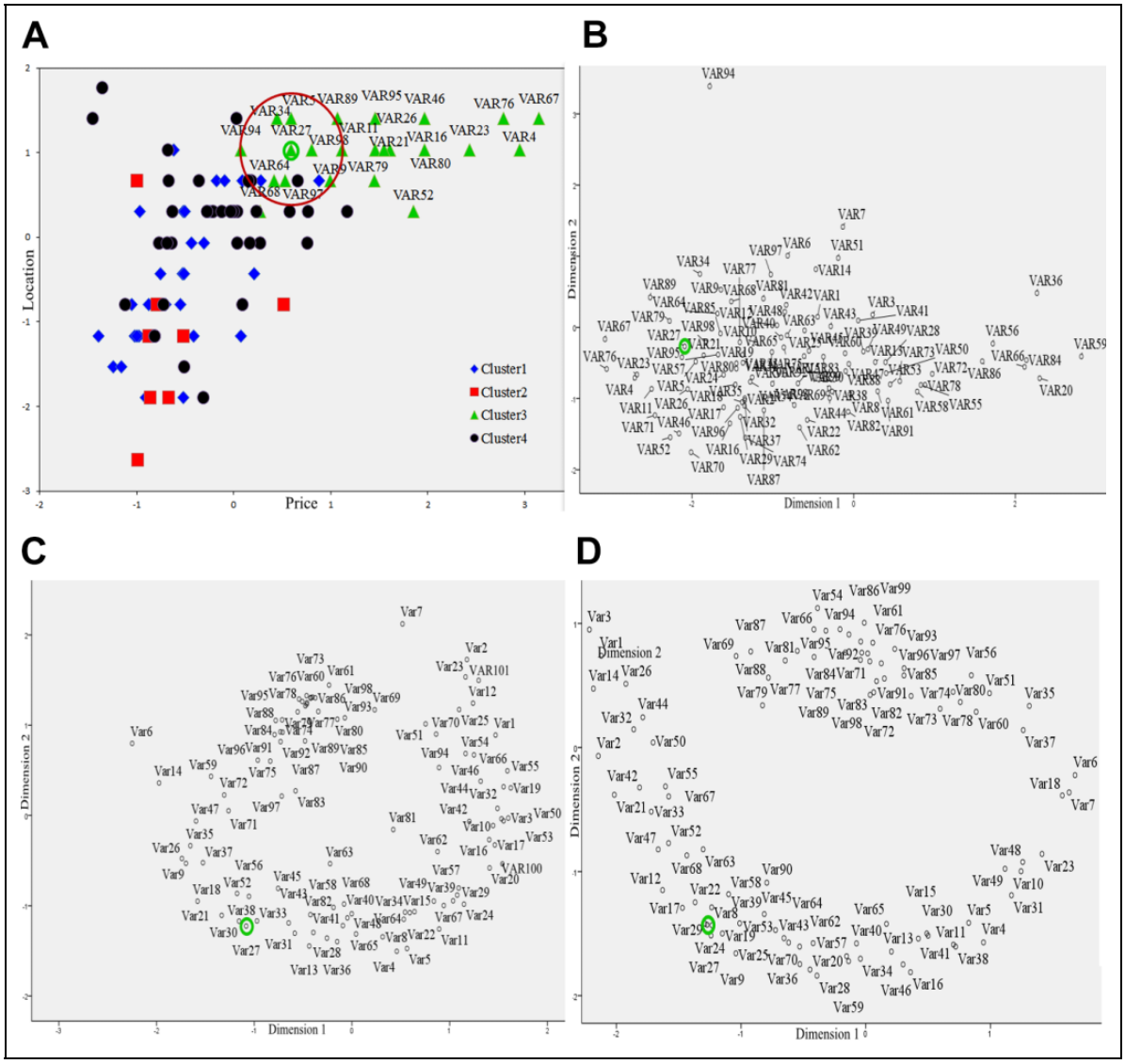

Figure E1. Visualization of competitive maps between the improved *k*NN model and MDS techniques. (A) The improved *k*NN using weighted Euclidean distance. (B) MDS using standard Euclidean distance. (C) MDS using cosine similarity. (D) MDS using the Pearson correlation coefficients.

Appendix F

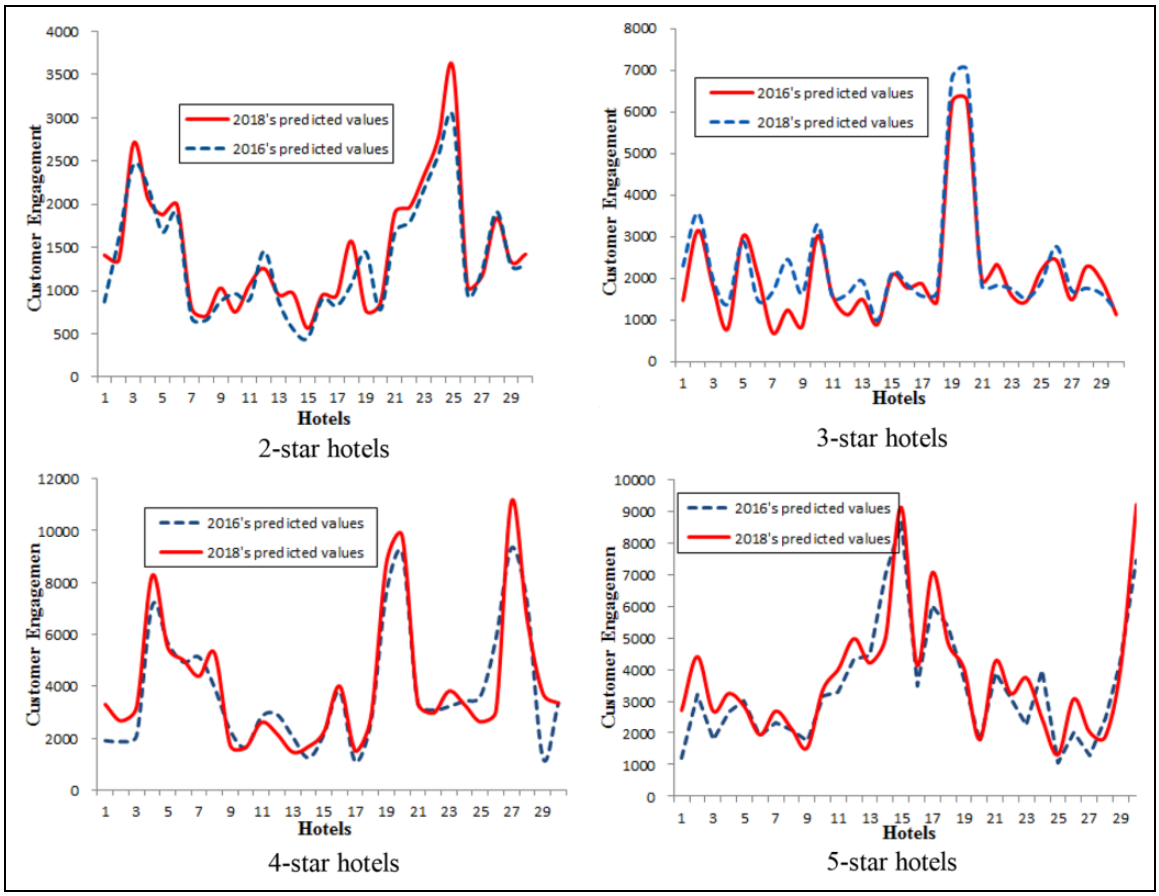

Figure F1. The trend charts of predicted values for different hotel star ratings in 2016 and 2018.

Appendix G

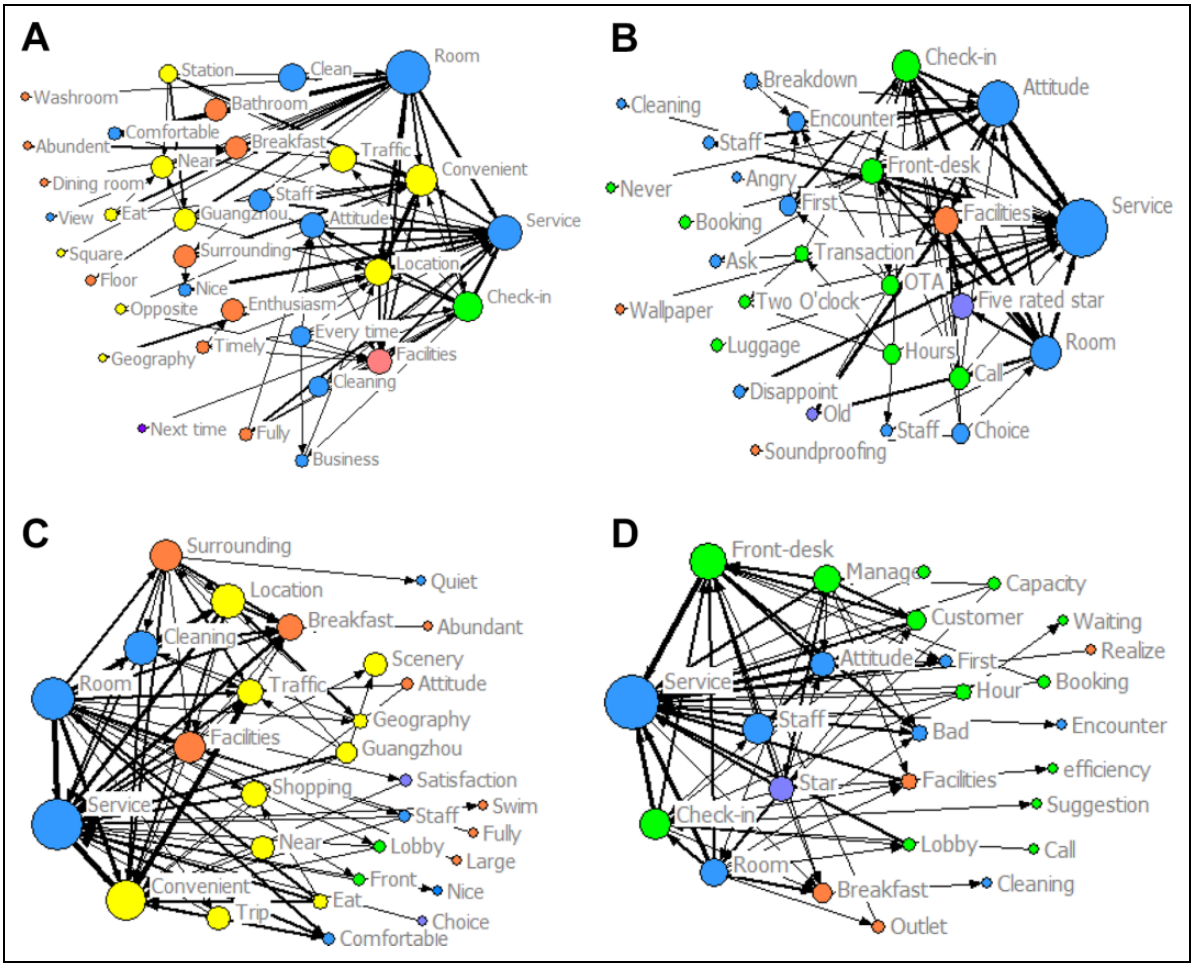

Figure G1. (A) Positive topics of Westin Guangzhou. (B) Negative topics of Westin Guangzhou. (C) Positive topics of Sofitel Guangzhou Sunrich Hotel. (D) Negative topics of Sofitel Guangzhou Sunrich Hotel.

Two Co-Occurrence Networks Generated by the LDA Topic Modeling Approach

Figures G1A and B show the positive and negative co-occurrence networks of the Westin Guangzhou Hotel. Figures G1C and D show the positive and negative co-occurrence networks of the Sofitel Guangzhou Sunrich Hotel. The key words regarding the topics of location, amenities, value, experience, and transaction are shown in yellow, orange, purple, blue, and green, respectively.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study is supported by National Natural Science Foundation of China (72071080, 71771090, 71471066), Natural Science Foundation of Guangdong Province (2019A1515010763, 2019A1515011768), Key Softscience Project of Guangdong Provice (2020B1010010001), Fundamental Research Funds for the Central Universities, SCUT (ZDPY201905, ZDPY201907) and the British Academy (SRG20\200985).

ORCID iD

Fei Ye https://orcid.org/0000-0002-0749-7393
Minhao Zhang https://orcid.org/0000-0002-1334-4481

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Please refer to Chinese Travel Consumer Report 2017–2018 at https://www.reutersevents.com/travel/distribution-strategies/chinese-travel-consumer-report-2017-2018
2. Please see the 2017 Ctrip Hotel White Paper at https://www.traveldaily.cn/images/201801/2017Ctrip_hotel_data.pdf. A summary of the report in English is available at https://www.chinatravelnews.com/article/119723

References

Algesheimer, R., S. Borle, U. M. Dholakia, and S. S. Singh (2011), "The Impact of Customer Community Participation on Customer Behaviors: An Empirical Investigation," *Marketing Science*, 29 (4), 756-769.

Amel, D. F. and S. A. Rhoades (1988), "Strategic Groups in Banking," *The Review of Economics and Statistics*, 70 (4), 685-689.

Antons, D. and C. F. Breidbach (2018), "Big Data, Big Insights? Advancing Service Innovation and Design with Machine Learning," *Journal of Service Research*, 21 (1), 17-39.

Arora, A., S. Bansal, C. Kandpal, R. Aswani, and Y. Dwivedi (2019), "Measuring Social Media Influencer Index-Insights from Facebook, Twitter and Instagram," *Journal of Retailing and Consumer Services*, 49 (3), 86-101.

Baum, J. A. C. and T. K. Lant (2003), "Hits and Misses: Managers' (mis)Categorization of Competitors in the Manhattan Hotel Industry," *Strategic Management*, 20 (6), 119-156.

Blei, David M. (2012), "Probabilistic Topic Models," *Communications of the ACM*, 55 (4), 77-84.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3 (4-5), 993-1022.

Brodie, R. J., L. D. Hollebeek, B. Jurić, and A. Ilić (2011), "Customer Engagement: Conceptual Domain, Fundamental Propositions, and Implications for Research," *Journal of Service Research*, 14 (3), 252-271.

Brown, J. R. and C. S. Dev (2000), "Improving Productivity in a Service Business: Evidence from the Hotel Industry," *Journal of Service Research*, 2 (4), 339-354.

Buja, A., D. F. Swayne, M. L. Littman, N. Dean, H. Hofmann, and L. Chen (2008), "Data Visualization with Multidimensional Scaling," *Journal of Computational and Graphical Statistics*, 17 (2), 444-472.

Carroll, J. Douglas and Phipps Arabie (1980), "Multidimensional Scaling," *Annual Review of Psychology*, 31 (1), 607-649.

Chen, Y. and S. Yao (2016), "Sequential Search with Refinement: Model and Application with Click-Stream Data," *Management Science*, 63 (12), 4345-4365.

Cheng, D., S. Zhang, Z. Deng, Y. Zhu, and M. Zong (2014), "kNN Algorithm with Data-Driven k Value," in *Advanced Data Mining and Applications*, X. Luo, J. X. Yu and Z. Li (eds). ADMA 2014. Cham, Switzerland: Lecture Notes in Computer Science, Springer, 8933 (12), 499-512.

Cheng, M. (2016), "Sharing Economy: A Review and Agenda for Future Research," *International Journal of Hospitality Management*, 57 (6), 60-70.

China National Tourism Bureau (2016), "China Tourism Statistic Yearbook," accessed October 10, 2017, [available at https://zwgk.mct.gov.cn/auto255/201710/t20171010_832494.html?keywords=]

Chinese Travel Consumer Report 2017-2018. (accessed October 5), [available at https://www.reutersevents.com/travel/distribution-strategies/chinese-travel-consumer-report-2017-2018].

Choi, S. C. (1991), "Price Competition in a Channel Structure with a Common Retailer," *Marketing Science*, 10 (4), 271-296.

Clark, B. H. and D. B. Montgomery (1999), "Managerial Identification of Competitors," *Journal of Marketing*, 63 (3), 67-83.

Cooper, L. G. and A. Inoue (1996), "Building Market Structures from Consumer Preferences," *Journal of Marketing Research*, 33 (3), 293-306.

Cooper, L. G. (1983), "A Review of Multidimensional Scaling in Marketing Research," *Applied Psychological Measurement*, 7 (4), 427-450.

DeSarbo, W. S. and R. Grewal (2007), "An Alternative Efficient Representation of Demand-Based Competitive Asymmetry," *Strategic Management Journal*, 28 (7), 755-766.

Diaconis, P., S. Goel, and S. Holmes (2008), "Horseshoes in Multidimensional Scaling and Local Kernel Methods," *The Annals of Applied Statistics*, 2 (3), 777-807.

Du, R. Y., Y. Hu, and S. Damangir (2015), "Leveraging Trends in Online Searches for Product Features in Market Response Modeling," *Journal of Marketing*, 79 (1), 29-43.

Filieri, R., F. McLeay, B. Tsui, and Z. Lin (2018), "Consumer Perceptions of Information Helpfulness and Determinants of Purchase Intention in Online Consumer Reviews of Services," *Information and Management*, 55 (8), 956-970.

Gao, S., O. Tang, H. Wang, and P. Yin (2018), "Identifying Competitors through Comparative Relation Mining of Online Reviews in the Restaurant Industry," *International Journal of Hospitality Management*, 71 (9), 19-32.

Golub, G. H. and H. G. Wahba (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21 (2), 215-223.

Gur, F. A. and T. Greckhamer (2019), "Know Thy Enemy: A Review and Agenda for Research on Competitor Identification," *Journal of Management*, 45 (5), 2072-2100.

Hartmann, J., J. Huppertz, C. Schamp, and M. Heitmann (2019), "Comparing Automated Text Classification Methods," *International Journal of Research in Marketing*, 36 (1), 20-38.

Hatzijordanou, N., N. Bohn, and O. Terzidis (2019), "A Systematic Literature Review on Competitor Analysis: Status Quo and Startup Specifics," *Management Review Quarterly*, 4 (3), 1-44.

Hoffman, M., F. R. Bach, and D. M. Blei (2010), "Online Learning for Latent Dirichlet Allocation," *Advances in Neural Information Processing Systems*, 23 (11), 856-864.

Holloway, B. B. and S. E. Beatty (2003), "Service Failure in Online retailing: A Recovery Opportunity," *Journal of Service Research*, 6 (1), 92-105. https://www.chinayearbooks.com/tags/the-yearbook-of-china-tourism-statistics

Jin, J., P. Ji, and R. Gu (2016), "Identifying Comparative Customer Requirements from Product Online Reviews for Competitor Analysis," *Engineering Applications of Artificial Intelligence*, 49 (3), 61-73.

Keiningham, T. L., A. Buoye, and J. Ball (2015), "Competitive Context Is Everything: Moving from Absolute to Relative Metrics," *Global Economics and Management Review*, 20 (2), 18-25.

Keiningham, T. L., B. Cooil, E. C. Malthouse, A. Buoye, L. Aksoy, A. D. Keyser, and B. Larivière (2015), "Perceptions Are Relative: An Examination of the Relationship between Relative Satisfaction Metrics and Share of Wallet," *Journal of Service Management*, 26 (1), 2-43.

Keiningham, T. L., S. Gupta, L. Aksoy, and A. Buoye (2014), "The High Price of Customer Satisfaction," *MIT Sloan Management Review*, 55 (3), 37.

Kim, J. B., P. Albuquerque, and B. J. Bronnenberg (2011), "Mapping Online Consumer Search," *Journal of Marketing Research*, 48 (1), 13-27.

Kim, J. Y. and L. Canina (2011), "Competitive Sets for Lodging Properties," *Cornell Hospitality Quarterly*, 52 (1), 20-34.

Kumar, V. and A. Pansari (2016), "Competitive Advantage through Engagement," *Journal of Marketing Research*, 53 (4), 497-514.

Kumar, V., L. Aksoy, B. Donkers, R. Venkatesan, T. Wiesel, and S. Tillmanns (2010), "Undervalued or Overvalued Customers: Capturing Total Customer Engagement Value," *Journal of Service Research*, 13 (3), 297-310.

Leung, D., R. Law, and A. L. Lee (2011), "The Perceived Destination Image of Hong Kong on Ctrip.com," *International Journal of Tourism Research*, 13 (2), 124-140.

Li, J. and S. Netessine (2012), "Who are my competitors? -Let the customer decide," INSEAD Working Paper No. 2012/84/TOM, [available at https://ssrn.com/abstract=2147638

Liu, Yong (2006), "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing*, 70 (7), 74-89.

Lu, A. C. C., D. Gursoy, and C. Y. R. Lu (2016), "Antecedents and Outcomes of Consumers' Confusion in the Online Tourism Domain," *Annals of Tourism Research*, 57 (2), 76-93.

Lu, K. and A. Elwalda (2016), "The Impact of Online Customer Reviews (OCRS) on Customers' Purchase Decisions: An Exploration of the Main Dimensions of OCRS," *Journal of Customer Behavior*, 15 (2), 123-152.

Mankad, S., H. S. Han, J. Goh, and S. Gavirneni (2016), "Understanding Online Hotel Reviews through Automated Text Analysis," *Service Science*, 8 (2), 124-138.

Mariani, M. M., M. Borghi, and U. Gretzel (2019), "Online Reviews: Differences by Submission Device," *Tourism Management*, 70 (8), 295-298.

Mathwick, C. and J. Mosteller (2017), "Online Reviewer Engagement: A Typology based on Reviewer Motivations," *Journal of Service Research*, 20 (2), 204-218.

Mitani, Y. and Y. Hamamoto (2006), "A Local Mean-Based Nonparametric Classifier," *Pattern Recognition Letters*, 27 (10), 1151-1159.

Modha, D. S. and W. S. Spangler (2003), "Feature Weighting in k-Means Clustering," *Machine Learning*, 52 (3), 217-237.

Moe, W. W. and M. Trusov (2011), "Measuring the Value of Social Dynamics in Online Product Ratings Forums," *Journal of Marketing Research*, 48 (3), 444-456.

Mohammed, I., B. D. Guillet, and R. Law (2014), "Competitor Set Identification in the Hotel Industry: A Case Study of a Full-Service Hotel in Hong Kong," *International Journal of Hospitality Management*, 39 (39), 29-40.

Moore, W. L. and M. B. Holbrook (1982), "On the Predictive Validity of Joint-Space Models in Consumer Evaluations of New Concepts," *Journal of Consumer Research*, 9 (2), 206-210.

Nam, H., Y. V. Joshi, and P. K. Kannan (2017), "Harvesting Brand Information from Social Tags," *Journal of Marketing*, 81 (4), 88-108.

Nasution, H. N. and F. T. Mavondo (2008), "Organisational Capabilities: Antecedents and Implications for Customer Value," *European Journal of Marketing*, 42 (3/4), 477-501.

Netzer, O., R. Feldman, J. Goldenberg, and M. Fresko (2012), "Mine Your Own Business: Market-Structure Surveillance through Text Mining," *Marketing Science*, 31 (3), 521-543.

Ng, D., R. Westgren, and S. Sonka (2009), "Competitive Blind Spots in an Institutional Field," *Strategic Management Journal*, 30 (4), 349-369.

Ordenes, F. V., B. Theodoulidis, J. Burton, T. Gruber, and M. Zaki (2014), "Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach," *Journal of Service Research*, 17 (3), 278-295.

Pan, B., L. Zhang, and R. Law (2013), "The Complex Matter of Online Hotel Choice," *Cornell Hosp. Q*, 54 (1), 74-83.

Parasuraman, A., L. L. Berry, and V. A. Zeithaml (1993), "More on Improving Service Quality Measurement," *Journal of Retailing*, 69 (1), 140-147.

Pavlou, P. A. and A. Dimoka (2006), "The Nature and Role of Feedback text Comments in online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation," *Information Systems Research*, 17 (4), 392-414.

Pelsmacker, P. D., S. V. Tilburg, and C. Holthof (2018), "Digital Marketing Strategies, Online Reviews and Hotel Performance," *International Journal of Hospitality Management*, 72 (1), 47-55.

Peteraf, M. A. and M. E. Bergen (2003), "Scanning Dynamic Competitive Landscapes: A Market-Based and Resource-Based Framework," *Strategic Management Journal*, 24 (10), 1027-1041.

Phillips, P., K. Zigan, M. M. S. Silva, and R. Schegg (2015), "The Interactive Effects of Online Reviews on the Determinants of Swiss Hotel Performance: A Neural Network Analysis," *Tourism Management*, 50 (2), 130-141.

Rapp, A., R. Agnihotri, T. L. Baker, and J. Andzulis (2015), "Competitive Intelligence Collection and Use by Sales and Service Representatives: How Managers' Recognition and Autonomy Moderate Individual Performance," *Journal of the Academy of Marketing Science*, 43 (3), 357-374.

Ringel, Daniel M. and Bernd Skiera (2016), "Visualizing Asymmetric Competition among More Than 1,000 Products Using Big Search Data," *Marketing Science*, 35 (3), 511-534.

Ritov, Y. (1990), "Estimation in a Linear Regression Model with Censored Data [J]," *Annals of Statistics*, 18 (1), 303-328.

Roos, I., B. Edvardsson, and A. Gustafsson (2004), "Customer Switching Patterns in Competitive and Noncompetitive Service Industries," *Journal of Service Research*, 6 (3), 256-271.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 5 (3), 3-55.

Shao, T. and M. Kenney (2018), *Ctrip: China's Online Travel Platform—Local Giant or Global Competitor?* Social Science Electronic Publishing.

Sidhu, J. S., E. J. Nijssen, and H. R. Commandeur (2000), "Business Domain Definition Practice: Does It Affect Organisational Performance? *Long Range Planning*, 33 (3), 376-401.

Sohn, M. H., T. You, and S. L. Lee (2003), "Corporate Strategies, Environmental Forces, and Performance Measures: A Weighting Decision Support System Using the k-Nearest Neighbor Technique," *Expert Systems With Applications*, 25 (3), 279-292.

Song, B. D. and Y. D. Ko (2017), "Quantitative Approaches for Location Decision Strategies of a Hotel Chain Network," *International Journal of Hospitality Management*, 67 (8), 75-86.

Tan, H., X. Lv, X. Liu, and D. Gursoy (2018), "Evaluation Nudge: Effect of Evaluation Mode of Online Customer Reviews on Consumers' Preferences," *Tourism Management*, 65 (9), 29-40.

Thakur, R. (2018), "Customer Engagement and Online Reviews," *Journal of Retailing and Consumer Services*, 41 (11), 48-59.

Tsai, H. H. and I. Y. Lu (2006), "The Evaluation of Service Quality Using Generalized Choquet Integral," *Information Sciences*, 176 (6), 640-663.

Wang, W., F. Yi, and W. Dai (2018), "Topic Analysis of Online Reviews for Two Competitive Products Using Latent Dirichlet Allocation," *Electronic Commerce Research and Applications*, 29 (4), 142-156.

Wang, Y., M. Zhang, Y. K. Tse, and H. K. Chan (2020), "Unpacking the Impact of Social Media Analytics on Customer Satisfaction: Do External Stakeholder Characteristics Matter? *International Journal of Operations & Production Management*, 40 (5), 647-669.

Weatherford, L. R. and S. E. Bodily (1992), "A Taxonomy and Research Overview of Perishable-Asset Revenue Management: Yield Management, Overbooking, and Pricing," *Operations Research*, 40 (5), 831-844.

Wieringa, J. E. and P. C. Verhoef (2007), "Understanding Customer Switching Behavior in a Liberalizing Service Market: An Exploratory Study," *Journal of Service Research*, 10 (2), 174-186.

World Travel Market (2014), *World Travel Market Global Trends Report* (accessed October 21, 2019), [available at http://www.wtmlondon.com/RXUK/RXUK_WTMLondon/2015/documents/WTM-Global-Trends-2014.pdf].

Wu, J. and P. Olk (2014), "Technological Advantage, Alliances with Customers, Local Knowledge and Competitor Identification," *Journal of Business Research*, 67 (10), 2106-2114.

Wu, L., A. S. Mattila, C. Wang, and L. Hanks (2016), "The Impact of Power on Service Customers' Willingness to Post Online Reviews," *Journal of Service Research*, 19 (2), 224-238.

Wu, X., V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg (2008), "Top 10 Algorithms in Data Mining," *Knowledge and Information Systems* 14 (12), 1-37.

Ye, Q., R. Law, and B. Gu (2009), "The Impact of Online User Reviews on Hotel Room Sales," *International Journal of Hospitality Management*, 28 (1), 180-182.

Zhan, Y., K. H. Tan, L. Chung, L. Chen, and X. Xing (2020), "Leveraging Social Media in New Product Development: Organisational Learning Processes, Mechanisms and Evidence from China," *International Journal of Operations & Production Management*, 40 (5), 671-695.

Author Biographies

Fei Ye is a professor in operations and supply chain management at South China University of Technology. His research interest includes supply chain management, services management, and sustainable operations management. He has published two books and more than 150 research articles in journals, including *Journal of Retailing*, *Naval Research Logistics*, *European Journal of Operational Research*, *International Journal of Operations & Production Management*, *International Journal of Production Economics*, and so on.

Qian Xia is a PhD student at the School of Business Administration, South China University of Technology. Her research interests are mainly focused on hotel supply chain management and big data analytics.

Minhao Zhang is a lecturer in operations management at the University of Bristol. His research interest includes managerial supply chain risk management, quality management, and the social media analytics. He is currently serving as an editorial member of *Enterprise Information Systems.* He has published in *International Journal of Operations & Production Management*, *Industrial Marketing Management*, *IEEE Transactions on Engineering Management*, *Supply Chain Management: An International Journal*, *Journal of*

Business Research, *International Journal of Production Economics*, and *R&D Management.*

Yuanzhu Zhan is a lecturer in operations and supply chain management at the University of Liverpool. His research interests are in the areas of operations management, big data and analytics, innovation management, and sustainable supply chain management. His research has been published in various journals, including the *International Journal of Operations and Production Management*, *European Journal of Operational Research*, *International Journal of Production Research*, *International Journal of Production Economics*, *Transportation Research Part E: Logistics and Transportation Review*, *Annals of Operations Research*, and *R&D Management*.

Yina Li is a professor in operations and supply chain management at South China University of Technology. Her research interest includes supply chain management, green innovation management, and supply chain quality management. She has published over 50 research articles in journals, including *Journal of Retailing*, *European Journal of Operational Research*, *International Journal of Production Research*, *International Journal of Operations & Production Management*, *Journal of Environmental Management*, *International Journal of Production Economics*, and so on.